MODELING BIOLOGICAL INTERACTIONS USING SUPERVISED MACHINE LEARNING

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Julie L. Yang

May 2016

MODELING BIOLOGICAL INTERACTIONS USING SUPERVISED MACHINE LEARNING

Julie L. Yang, Ph.D.

Cornell University 2016

Oncoprotein-targeted drug therapies offer an alternative to cytotoxic drugs for the treatment of cancer [11, 20, 23, 29]. However, resistance to targeted therapies poses a major clinical challenge to their broader use. In addition to acquired resistance, where cancer cells acquire mutations under the selective pressure of drug treatment, recent studies have implicated cells of the tumor microenvironment in mediating innate resistance to targeted therapies [41, 48]. Stromal cells can confer resistance by secreting cytokines that activate alternative signaling pathways in cancer cells so that they can continue to grow and proliferate despite exposure to drug. In this dissertation, I use a supervised machine learning approach to model tumor-stromal interactions that mediate drug resistance from a published data set where stromal-mediated drug resistance is measured from co-culture experiments. My model, a multi-task bilinear regression, called multi-task affinity regression predicts how cytokines secreted by stromal cells interact with pathways in cancer cells to mediate innate drug resistance to molecularly targeted therapies. I computationally identified and experimentally validated HGF as a secreted factor that mediates resistance to the EGFR inhibitors in lung cancer cell lines. I also compared my model to an alternative method in the classification setting called multi-task pairwise support vector machine (SVM). The method was also used to model binding of DNA/RNA to transcription factors (TFs)/RNA binding proteins (RBPs) in protein binding

microarray and RNAcompete experiments [40]. We hope this work will provide new insights into how stromal cells promote drug resistance and how we might treat cancer with combination therapies that target the tumor microenvironment. More generally this dissertation can serve as a point of reference for scientists looking to use supervised machine learning methods to model biological interactions where a high-throughput affinity readout is available.

BIOGRAPHICAL SKETCH

Julie L. Yang was born in Shanghai, China. She immigrated with her family to the United States. She grew up in New York City where she completed her high school education at Stuyvesant High School. She attended college at Carnegie Mellon University where she earned her Bachelor of Science degree in Mathematics with a minor in Physics. Julie joined the Tri-Institutional Training Program in Computational Biology and Medicine in the summer of 2010 and subsequently joined the Leslie Laboratory in the fall of 2011 to pursue her dissertation research. Julie's dissertation research focused on applying supervised machine learning methods to model tumor-stromal interactions that mediate innate resistance against targeted cancer therapies.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

I would like to thank all of those people who helped make this dissertation possible. Firstly I would like to thank my advisor Dr. Christina Leslie for mentoring me through my graduate studies. I have learned about highly interesting applications of machine learning to high throughput biological and genomic data sets. I truly appreciate her patience, guidance, and support throughout the years I have worked with her. Under her guidance, I have learned that I really enjoy the work that I do. I want to give special thanks go to my special committee members for agreeing to serve on my committee: Dr. Olivier Elemento, Dr. Johanna Joyce, and Dr. Haiyuan Yu. Thank you for your support and insightful comments and counsel throughout this process.

I would like to thank the Tri-Institutional Training Program in Computational Biology and Medicine (CBM) for the opportunity to pursue my doctoral training in this wonderful vibrant research community. I would like to thank Dr. David Christini. He is an excellent advisor and stands as a example to us graduate students. Thanks to Ms. Kathleen Pickering and Ms. Margie Hinonangan-Mendoza for their support throughout my time in graduate school. I have been fortunate to learn from many exceptional scientists through collaboration across research groups and institutions. I want to thank my collaborators: Dr. Johanna Joyce and Dr. Bobby Bowman with whom I have worked closely with on the work presented in this dissertation.

I want to thank my colleagues from the Leslie lab: Raphael Pelossof, Alvaro J. Gonzlez, Hatice Ulku Osmanbeyoglu, Yuri Pritykin, Lee Zamparo, Yi Zhong, Meghana Kshirsagar, Steve Lianoglou, Manu Setty, Mark Carty, Yuheng Lu, Irtisha Singh, Lauren Fairchild, Hyunwoo Cho, Alexander Perez, Han Yuan, Sagar Chhangawala, and Angela Yu. You friendship has been an indescribable

blessing in my life. I would like to acknowledge my fellow graduate students from the CBM program. It has been wonderful to be your fellow graduate student in this graduate program.

Lastly, but in no way least, I would like to thank my family and friends for their love and support throughout this journey. I want to thank my mother, Yun Lu, as she pursued her Masters and Ph.D. degree in computer science as well and has served as the direct influence for me to apply to a computational biology program. I really admire her for her excellent work ethic as an electron microscopist doing excellent research work in neuroscience in C.elegans. I want to thank my dad, Ming Yang, who is a biophysicist as he has encouraged and always wanted me to pursue research in cancer biology and to commemorate my grandmother- my dad's mom, as she died of stomach cancer in China. I want to thank my older brother Charles Yang, his wife Jennifer, and my niece Sophia for their love for each other as a family. I want to thank my grandmother Shui Ping Zhang- my mom's mom as she taught me mathematics when I was in elementary school and she is, for a large part, the reason that I have the mathematical skill set necessary to pursue a PhD. in Computational Biology. I want to thank the rest of my dear family: my Aunt Ningna Lu, her husband Jin Juing Wang and my cousin Yunfan Wang and my Uncle Dong Ming Lu.

I want to thank God and His son, Jesus Christ. Thank You for Your abundant blessings in my life. May You be glorified. I want to thank my church family the Agape Churches and my church family in New York, Ecclesia New York. Thank you for loving me, partnering with me, daily being involved in my life, and enabling me to fully live out my life in God's victory.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vii
List of Figures . . . x

1 Introduction **1**
1.1 Background . . . 1
1.2 Summary outline . . . 5
1.2.1 Statistical data analysis of tumor-stromal co-culture drug screen . . . 6
1.2.2 Modeling tumor-stromal interactions that mediate innate resistance to cancer therapies . . . 7
1.2.3 Application of affinity regression to PBM and RNAcompete data and comparison to multi-task pairwise SVM . . . 9

2 Statistical data analysis of tumor-stromal co-culture drug screen **11**
2.1 Preliminary studies . . . 11
2.1.1 Mean drug rescue score in co-cultures treated with different drug dosages . . . 12
2.1.2 Dosage comparison of stromal mediated rescue in melanoma and colorectal co-cultures treated with PLX4720 13
2.2 Correlation Analysis . . . 15
2.2.1 Reanalysis of the correlation in cytokine arrays with average rescue score of melanoma cell lines treated with PLX4720 . . . 15
2.2.2 Analysis of the correlation in cytokine arrays with rescue scores in individual melanoma cell lines treated with PLX4720 . . . 18
2.2.3 Correlation analysis for cancer subtypes across co-cultures treated with anticancer drugs . . . 19

3 Modeling tumor-stromal interactions that mediate innate resistance to targeted cancer therapies **34**
3.1 Introduction . . . 34
3.1.1 Background . . . 34
3.1.2 Prior Work . . . 36
3.2 Results . . . 40
3.2.1 Multi-task affinity regression models tumor-stromal interactions in co-culture drug treatment experiments . . . 40
3.2.2 Multi-task affinity regression outperforms nearest neighbor methods for predicting rescue in co-culture experiments 45

3.2.3 Multi-task affinity regression recovers HGF as a stromal factor that rescues melanoma cells treated with PLX4720 . 46
3.2.4 Affinity regression implicates MET, MYC and PI3K pathways in stromal cell-mediated PLX4720 resistance 49
3.2.5 Multi-task training learns a pan-EGFR inhibitor model of stromal-mediated resistance 52
3.2.6 Experimental validation confirms HGF as a novel stromal factor mediating resistance to afatinib and erlotinib in non-small cell lung cancer cells 55
3.3 Methods 56
3.3.1 Rescue score calculation 56
3.3.2 Cytokine array processing 58
3.3.3 Pathway scores 58
3.3.4 Multi-task affinity regression 59
3.3.5 Parameter optimization 60
3.3.6 Comparison of multi-task with single-task affinity regression 61
3.3.7 Comparison of multi-task affinity regression with nearest neighbor methods 61
3.3.8 Empirical null models 62
3.3.9 Cells, inhibitors, and cytokines 63
3.4 Conclusion 64
3.5 Comparison of multi-task affinity regression to multi-task pairwise SVM 65
3.5.1 Optimization Problem for Pairwise SVM 65
3.5.2 Primal Optimization Problem for Multi-task Pairwise SVM 69
3.5.3 Dual Optimization Problem for Multi-task Pairwise SVM . 70
3.6 Performance Comparison of multi-task affinity regression to multi-task pairwise SVM 71

4 Application of affinity regression to PBM and RNAcompete data 77
4.1 Application of affinity regression to PBM and RNAcompete data 77
4.1.1 Affinity regression modeling of TF/RBP and DNA/RNA interactions in PBM and RNAcompete data 77
4.1.2 Example of predicted Z-scores 79
4.1.3 Motif visualization and comparison of affinity regression motifs with nearest neighbor motifs 79
4.1.4 Comparison of Kullback-Leibler divergence of affinity regression motifs with nearest neighbor motifs 81

5 Conclusion 86
5.1 Conclusion and future directions 86
5.1.1 Introduction 86
5.1.2 Stromal-mediated drug resistance 87

5.1.3 Learning families of transcription factors and RNA binding proteins from PBM and RNAcompete data 90
5.1.4 Conclusion 91

Bibliography **92**

LIST OF FIGURES

2.1 **Rescue of colorectal cancer cell lines from gemcitabine (Gem) by HDF stromal cells.** HT-29 was treated with 0.02 μ M Gemcitabine. Fluorescence microscopy looking at GFP positive cancer cells at day 7. All 4 quadruplicate wells are shown. 12

2.2 **Mean drug rescue score averaged across all tumor-stromal co-cultures indexed by the drug dosage.** The average rescue score for co-cultures treated was calculated and plotted for each drug dosage. 14

2.3 **Relative proliferation with stroma versus relative proliferation without stroma in melanoma cancer cell lines.** The relative proliferation of the cancer cells in co-culture with melanoma cell lines was plotted against the relative proliferation of cancer cells in monoculture for cancer cell lines treated with vemurafenib. . 16

2.4 **Relative Proliferation with stroma versus relative proliferation without stroma in colorectal cancer cell lines.** The relative proliferation of cancer cells in co-culture with colorectal cell lines was plotted against the relative proliferation of cancer cells in monoculture for cancer cell lines treated with vemurafenib. . . . 17

2.5 **Correlations for each of the cytokines in cytokine array one with stromal average melanoma rescue score R_M.** The correlation of the secretion of each cytokine in cytokine array one with the stromal average rescue score for melanoma cell lines treated with PLX4720 was calculated. The stromal average rescue score for melanoma was calculated for each of the stromal cell lines as the average rescue score across all melanoma cell lines in co-culture with the stromal cell line. 21

2.6 **Correlations for each of the cytokines in cytokine array two with stromal average melanoma rescue score R_M.** The correlation of the secretion of each cytokine in cytokine array one with the stromal average rescue score for melanoma cell lines treated with PLX4720 was calculated. The stromal average rescue score for melanoma was calculated for each of the stromal cell lines as the average rescue score across all melanoma cell lines in co-culture with the stromal cell line. 22

2.7 **Box plots of correlations of cytokines secretion in cytokine array one with rescue of melanoma cell lines treated with PLX4720 1 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array one with the rescue scores from the melanoma cell lines under the drug PLX4720 1 μM. 23

2.8 **Box plots of correlations of cytokines secretion in cytokine array two with rescue of melanoma cell lines treated with PLX4720 1 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array two with the rescue scores from the melanoma cell lines under the drug PLX4720 1 μM. . . . 24
2.9 **Box plots of correlations of cytokines secretion in cytokine array one with rescue of melanoma cell lines treated with PLX4720 2 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array one with the rescue scores from the melanoma cell lines under the drug PLX4720 2 μM. 25
2.10 **Box plots of correlations of cytokines secretion in cytokine array two with rescue of melanoma cell lines treated with PLX4720 2 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array two with the rescue scores from the melanoma cell lines under the drug PLX4720 2 μM. . . . 26
2.11 **Box plots of rescue scores plotted against GM-CSF secretion in SB590885 4 μM treated melanoma** The rescue scores are plotted as a function of the secretion of GM-CSF secretion in melanoma cell lines treated with SB590885 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines for the cancer subtypes is calculated. 27
2.12 **Box plots of rescue scores plotted against HGF secretion in vemurafenib 4 μM treated melanoma** The rescue scores are plotted as a function of the secretion of HGF secretion in melanoma cell lines treated with PLX4720 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 28
2.13 **Box plots of rescue scores plotted against HGF secretion in vemurafenib 2 μM treated melanoma** The rescue scores are plotted as a function of the secretion of HGF secretion in melanoma cell lines treated with PLX4720 2 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 29
2.14 **Box plots of rescue scores plotted against MMP-10 secretion in vemurafenib 1 μM treated melanoma** The rescue scores are plotted as a function of the secretion of MMP-10 secretion in melanoma cell lines treated with PLX4720 1 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 30

2.15 **Box plots of rescue scores plotted against VEGF-C secretion in afatinib 0.05 μM treated breast cancer** The rescue scores are plotted as a function of the secretion of VEGF-C secretion in breast cancer cell lines treated with afatinib 0.05 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 31

2.16 **Box plots of rescue scores plotted against MMP-10 secretion in lapatinib 4 μM treated head and neck cancer** The rescue scores are plotted as a function of the secretion of MMP-10 secretion in head and neck cancer cell lines treated with lapatinib 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 32

2.17 **Box plots of rescue scores plotted against HGF secretion in afatinib 0.03 μM treated non-small cell lung cancer** The rescue scores are plotted as a function of the secretion of HGF secretion in non-small cell lung cancer cell lines treated with afatinib 0.03 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown. 33

3.1 **Hierarchical clustering of the pathway scores groups cancer cell lines by tissue of origin** (a) Hierarchical clustering of the pathway scores for 75 curated pathways across 1036 cell lines from CCLE groups cancer cell lines by tissue of origin and clusters together growth factor signaling pathways (red) and cytokine/chemokine mediated pathways (blue). Hematopoietic, stomach, and large intestine cancer cell lines have high growth factor signaling pathway scores, while lung and skin cell lines have high cytokine and chemokine mediated pathway scores. (b) Her2 amplified breast cancer cell lines have higher ERBB2 signaling pathway scores than Her2 wild type breast cancer cell lines. (c) Melanoma cell lines with BRAF mutations have higher RAS/RAF signaling pathway scores than those with wildtype BRAF. 41

3.2 **Affinity regression predicts stromal-mediated rescue from targeted therapies from cancer cell line pathway scores and stromal cell cytokine data.** (a) Drug co-culture experiments quantify the extent of stromal-mediated rescue of cancer cells treated with targeted agents, assigning a rescue score to each (cancer cell line, stromal cell line, drug/dosage) combination. (b) Multi-task affinity regression learns to predict rescue scores from pathway score features of cancer cells and secreted cytokine levels of stromal cells using a regularized bilinear regression strategy. Each task trains on the co-culture experiments for a specific drug and dosage. The model is represented as an interaction matrix W_t between pathways and cytokines; S and C represent the feature matrices of cytokine expression values for stromal cells and pathway scores for cancer cells, respectively. Tasks corresponding to different dosages of the same drug or to drugs in the same class are jointly trained by shrinking model matrices W_t for different tasks towards the average task W_o. 43

3.3 **Cytokine and pathway mappings** (a) Multiplying a cancer cell lines pathway feature vector by the trained model W yields a vector of cytokine mapping scores. Large positive mapping scores identify cytokines predicted to mediate innate drug resistance when they interact with the cancer cell line. (b) Multiplying a stromal cell lines cytokine feature vectors by the trained model W yields a vector of pathway mapping scores. Large mapping scores identify cancer pathways whose dysregulation is predicted to mediate resistance or sensitivity in the presence of the stromal cell line. 44

3.4 **Multi-task affinity regression outperforms nearest-neighbor methods in 10-fold cross-validation experiments.** Models were trained in multi-task fashion for 54 drug tasks corresponding to multiple dosage levels of 13 anti-cancer therapeutics; pan-EGFR inhibitor and pan-BRAF inhibitor models were also trained. Multi-task affinity regression strongly outperformed (a) cancer nearest neighbor (P < 2.69 × 10-14, Wilcoxon signed rank test) and (b) stromal nearest neighbor (P < 4.77 × 10-13, Wilcoxon signed rank test). 46

3.5 **Multi-task affinity regression identifies HGF as the main cytokine that elicits resistance in PLX4720 treated melanoma cell lines.** (a) The heatmap of the cytokine mapping scores for melanoma and colorectal cell lines using the vemurafenib model clusters cancer cell lines by subtype. Orange boxes indicate a cytokine-cancer cell line pair significant mapping scores relative to an empirical null model (FDR < 5%, see Methods) and dark orange boxes indicate a cytokine-cancer cell line pair significant mapping scores relative to an empirical null model (FDR < 10%, see Methods), identifying HGF as the main cytokine that elicits resistance in melanoma cell lines. (b) For each cytokine, the bar plots show the number of melanoma cell lines (magenta) for which the mapping score attained significance of (FDR < 10%). This analysis suggests that stromal-secreted HGF and TNF-beta frequently mediate resistance to vemurafenib in lung cancer cells. 48

3.6 **Multi-task affinity regression implicates the MYC pathway in drug resistance and P53 regulation pathway in drug sensitivity in PLX4720 treated cell lines.** The heatmap of the pathway mapping scores for breast, skin, and lung fibroblasts using the vemurafenib model clusters stromal cell lines by their tissue of origin. Dark orange boxes indicate pairs whose mapping scores are significantly associated with drug resistance and dark blue boxes indicate pairs whose mapping scores are significantly associated with drug sensitivity (FDR <10%). This analysis implicates the MYC pathway in drug resistance and the P53 regulation pathway in drug sensitivity mediated by lung and breast fibroblasts. 50

3.7 **Multi-task affinity regression identifies the HGF and MET pathway and HGF and PI3K and PI3K/AKT pathways in drug resistance against vemurafenib** The heatmap of absolute values of cytokine-pathway interactions in the model matrix W. Orange boxes indicate cytokine-pathway interactions that are significant relative to an empirical null model (FDR < 5%, see Methods) and dark orange boxes indicate (FDR < 10%, see Methods) and identify (HGF, MET pathway) and (HGF, PI3K pathway) and (HGF, PI3K/AKT pathway) as significant interactions. 51

3.8 **Multi-task affinity regression identifies the HGF in drug resistance in afatnib treated co-cultures** (a) The heatmap shows cytokine mapping scores for breast cancer and non-small cell lung cancer (NSCLC) cells lines for the afatinib affinity regression model. Orange boxes indicate cytokine-cancer cell line pairs that are significantly associated with drug resistance relative to an empirical null model (FDR < 5%), while dark orange boxes indicate (FDR < 10%). (b) For each cytokine, the bar plots show the number of breast cancer cell lines (red) and the number of lung cancer cell lines (black) for which the mapping score attained significance. This analysis suggests that stromal-secreted HGF and IL-8 frequently mediate resistance to afatinib in lung cancer cells, but seldom in breast cancer cells. 53

3.9 **Multi-task affinity regression identifies the HGF in drug resistance in pan-EGFR inhibitor model** (a) The heatmap shows cytokine mapping scores for breast cancer, NSCLC, and head and neck squamous cell carcinoma (HNSCC) cell lines for the pan-EGFR affinity regression model with multi-task training across gefitinib, afatinib, erlotinib, and CL-387785 (b) For each cytokine, the bar plots show the number of breast cancer cell lines (red) and the number of lung cancer cell lines (black) for which the mapping score attained significance for the pan-EGFR affinity regression model. The analysis suggests that HGF and IL-8 frequently mediate resistance to EGFR inhibitors in lung cancer cells, while MMP-1 mediates resistance in breast cancer cells. . . 54

3.10 **Summary of HGF mediated drug resistance non-small cell lung cancer cell lines** Summary of drug resistance experiments comparing cancer cell line proliferation under treatment with two EGFR inhibitors, afatinib and erlotinib, across 3 NSCLC cell lines with and without HGF. The P value indicates the most significant increase in cancer cell counts over the drug dose response experiments in drug+HGF versus drug only conditions. 56

3.11 **Dose response curves and relative abundance of cancer proliferation at a specific drug dose for HGF mediated drug resistance in non-small cell lung cancer cell lines** (b,d,f,h) Dose response curves for drug+HGF (red) versus drug only (blue) for specific NSCLC cell lines, plotting relative abundance of cancer cells as a function of drug dose with HGF concentration fixed. Experiments are performed in triplicate. (c,e,g,i) Relative abundance of cancer cells at a specific drug dose (as shown), showing baseline, HGF only, drug only, and HGF values. 57

3.12 **Comparison of multi-task affinity regression with multi-task pairwise SVM in PLX4720 drug resistance** Multi-task affinity regression (AUC = 0.93) outperforms multi-task pairwise SVM (AUC = 0.89). Models were assessed using 10-fold cross-validation. 72
3.13 **Comparison of multi-task affinity regression with multi-task pairwise SVM in afatinib drug resistance** Multi-task affinity regression (AUC = 0.94) outperforms multi-task pairwise SVM (AUC = 0.72). Models were assessed using 10-fold cross-validation. 73
3.14 **Comparison of multi-task affinity regression with multi-task pairwise SVM in erlotinib drug resistance** Multi-task affinity regression (AUC = 0.95) outperforms multi-task pairwise SVM (AUC = 0.85). Models were assessed using 10-fold cross-validation. 74
3.15 **Comparison of multi-task affinity regression with multi-task pairwise SVM in gefitinib drug resistance** Multi-task affinity regression (AUC = 0.91) outperforms multi-task pairwise SVM (AUC = 0.86). Models were assessed using 10-fold cross-validation. 75
3.16 **Comparison of multi-task affinity regression with multi-task pairwise SVM AUC performance in 13 drugs and 2 drug groups.** Multi-task affinity regression outperforms multi-task pairwise SVM in AUC ($P < 3.05 \times 10\text{-}4$, Wilcoxon signed rank test). Models were assessed using 10-fold cross-validation. . . . 76

4.1 **Example of homeodomain predicted versus experimental Z-scores.** Example of predicted Z-scores from the Z-score affinity regression model, trained on 75 non-redundant mouse homeodomains, versus experimental Z-scores for SNAPOd2T00005194001, one of the diverse homeodomains assayed by Weirauch et al. [15] Binding motifs generated by PWM-Align-Z based on the top 100 8-mers predicted by affinity regression and the top 100 8-mers based on actual Z-scores are shown. . . . 80
4.2 **AR-derived motif prediction for RBPs.** AR motifs are generated by running PWM-Align-Z on the top 100 7-mers as predicted for held-out RBPs using the Z-score affinity regression model (10-fold cross-validation). In the inner circle, we show the ground truth motif obtained from the experimental data Y, the middle circle shows motif obtained by AR, and the outer circle shows the motif obtained by NN. Plotted are predictions for both RRM and KH-I domains. The RRM motifs are well predicted by both AR and NN; KH family proteins are less well represented in the data set, and KH-I motifs are harder to predict for both methods. 82

4.3 **RBP motif accuracy.** Predicted motifs were assessed for quality relative to ground truth by D_{KL}, computed by sliding the predicted PWM over the ground truth PWM and reporting the minimum D_{KL}. The NN (y-axis) and AR (x-axis) $\log(D_{KL})$ scores are plotted after subtracting the minimum $\log(D_{KL})$ for the data set. Motifs falling in the gray area satisfy a quality threshold equal to the median divergence between motifs from experimental replicates. AR-predicted motifs are significantly more accurate than NN motifs ($p < 7.66 \times 10\text{-}10$, Wilcoxon signed rank test). 83
4.4 **Probability density function of $\log(D_{KL})$ for affinity regression and nearest neighbor.** Probability density function of held-out $\log(D_{KL})$ for AR and NN. 84
4.5 **Cumulative distribution function of $\log(D_{KL})$ for affinity regression and nearest neighbor.** Cumulative distribution function of held-out $\log(D_{KL})$ for AR and NN. 85

CHAPTER 1

INTRODUCTION

1.1 Background

Targeted cancer therapies, which include small molecules and blocking antibodies that inhibit a specific molecular target within the tumor cell, have seen marked success in clinical trials in the past two decades [11, 20, 23, 29]. Targeted therapies represent a significant advance in personalized medicine as they are used specifically on patients whose tumors harbor the genetic mutations that give rise to targeted oncogenic proteins. As opposed to chemotherapeutic drugs, which kill rapidly dividing cells indiscriminantly, targeted therapies offer a promising treatment option that effects only cancer cells expressing specific molecules. The success of targeted therapies and their advantages over chemotherapy points to target therapies as a important alternative treatment against cancer. However, the emergence of drug resistance in the 2000s in clinical trials posed a serious challenge to the success of these treatments [1, 7, 22, 39]. In these clinical trials, a large percentage of the treatment cohort were either non-responders or responders who would fail to respond after a period of time. Drug resistance poses a major problem for targeted therapies, and it is important to study mechanisms of drug resistance to develop treatment strategies that will overcome these challenges. This thesis studies tumor microenvironment-mediated innate drug resistance by modeling molecular interactions between the tumor and stromal cells using supervised machine learning methods. This analysis provides new insights into the mechanisms that mediate drug resistance and may eventually suggest treatment strategies such as

combination therapies that can target both the tumor and microenvironment to overcome drug resistance.

Research on targeted therapies for cancer treatment can be traced back to as early as 1960 when researchers found a chromosomal abnormality leading to the expression of mutant kinase BCR-ABL on chromosome 22 in chronic granulocytic leukemia [29]. They named the chromosomal abnormality the Philadelphia Chromosome. Later imatinib (Gleevec) one of the first targeted cancer therapies was developed as an inhibitor of mutant kinase BCR-ABL. In the late 1990s and early 2000s, the FDA began to approve the first molecularly targeted cancer drugs. In 1997, the FDA approved rituximab (Rituxan) to treat patients with non-Hodgkin lymphoma [11]. In 1998, the FDA approved trastuzumab (Herceptin), a monoclonal antibody that was added to chemotherapy to treat women with advanced breast cancer that over-expressed HER2 [23]. In 2001, the FDA approved imatinib (Gleevec) to treat chronic myelogenous leukemia (CML) [20]. These targeted therapies allowed specific cohorts of patients to avoid the negative effects of cytotoxicity from chemotherapy.

Despite the initial success of targeted therapies, drug resistance soon emerged as a serious a challenge. There are two types of drug resistance: acquired resistance, which develops during the course of treatment in response to therapy, and innate resistance, which is inherent in the body and present even before treatment begins [19]. For many targeted therapies, some patients do not respond (innate resistance) or some patients fail to continue to respond after a period of time despite a significant initial response to therapy (acquired resistance). For imatinib (Gleevec), in the treatment of chronic myelogenous leukemia (CML), 33% of patients will have an inferior response, either fail-

ing to respond to primary therapy or demonstrating relapse after an initial response due to acquired drug resistance [7]. Trastuzumab (Herceptin) is an effective targeted drug in the treatment of HER2-positive breast cancer. However, 20% of early stage breast cancer patients and approximately 70% of patients with metastatic disease are resistant to treatment [28]. More recently imatinib (Gleevec) was also shown to block c-KIT tyrosine kinase in patients with KIT mutant melanoma [22]. Despite the initial response to imatinib treatment, drug resistance emerged for most cases after a short period of time eventually leading to relapse. Vemurafenib, a drug that targets $BRAF^{V600E}$ mutations, showed more than 50% response rates in patients carrying the BRAF mutation. Despite the high initial rate of response to therapy, a majority of the patients develop resistance to vemurafenib after approximately six months of treatment [39, 33]. Drug resistance poses a challenge to the success of targeted therapies, and it is important to study mechanisms of drug resistance in order to improve the duration of response.

Cancers develop due to the accumulation of genetic and epigenetic alterations that occur in initially normal cells. At same time, cancer cells develop within a host microenvironment that is composed of a heterogeneous population of stromal cells (fibroblasts, endothelial cells, adipocytes, immune cells and bone marrow-derived stem cells), stromal cell secreted factors, and the extracellular matrix. The host microenvironment has long been known to contribute to tumor initiation, progression and metastasis [26]. Recently researchers have proposed that the tumor microenvironment may also play a key role in modulating cancer drug efficacy. In 2012, Todd Golub's group at the Broad Institute [41] and Jeffrey Settleman's group at genentech [48] separately and concurrently showed that stromal cells from the tumor microenvironment confer drug resis-

tance through secreted factors. Targeted therapies were especially susceptible to stromal mediated drug resistance. It was shown that stromal cells of the tumor microenvironment could induce resistance in cancer cells by releasing soluble factors called cytokines that elicit drug resistance in tumor cells.

In my thesis research, I modeled tumor-stromal interactions that mediate drug resistance against anti-cancer targeted therapies. I used a multi-task bilinear regression method called multi-task affinity regression to model tumor-stromal interactions that mediate drug resistance, jointly training across co-cultures treated with different drugs of the same class. In my model, cytokines of the stromal cells are thought to signal to the tumor cells reactivating oncogenic signaling pathways in the tumor cells, leading to drug resistance. Through statistical analysis of the affinity regression models, I identified cell-cell signaling and cancer pathways involved in stromal-mediated drug resistance. Eventually, these approaches may lead to effective cancer treatment strategies such as combination drug therapies that target both the tumor and the microenvironment to counteract drug resistance.

Affinity regression is a bilinear regression where the observed data can be modeled as interactions between two kinds of inputs [40]. The algorithm learns a weighting on all such interactions that best explains the affinity of one input for the other given the observed data. Affinity regression is closely related to Partial Least Squares regression, developed by Herman Wold in 1966 [18], and to Canonical Correlation Analysis, developed by Harold Hotelling in 1936 [17]. Bilinear regression was described by Ruben Gabriel Kramer in 1995 [32]. Our bilinear regression method called affinity regression was developed by Raphael Pelossof in 2010. It involves singular value decomposition (SVD) compressions

of one of both inputs and regressing on the matrix of pairwise similarities of the output response variables. Here we formulate a multi-task version of this bilinear regression to analyze biological interactions of co-cultures that have been treated with multiple drug dosages. Our formulation is a bilinear regression without the SVD compressions and using the original output matrix rather than the kernelized version. We anticipate that our approach can be naturally applied to predict many other kinds of biological interactions where there is a high-throughput affinity read-out.

1.2 Summary outline

This thesis is divided into three chapters that cover the key contributions of my graduate work. Chapter 2 details the preliminary statistical data analysis I performed on the tumor-stromal co-culture drug screen data set obtained from Straussman et al. [41]. It includes the initial analyses of the drug resistance observed in the co-cultures and a correlation analysis of the cytokines responsible for mediating drug resistance in different subtypes of cancer treated with different anti-cancer therapies. Chapter 3 details the work I did to model tumor-stromal interactions that mediate innate drug resistance to anti-cancer therapies using multi-task affinity regression. I then explore an alternative method for modeling the pairwise molecular interactions between tumor cells and stromal cells using multi-task pairwise support vector machine (SVM) in a classification framework and compare the performance of this alternative method with multi-task affinity regression. Finally, chapter 4 describe another application of affinity regression to Protein Binding Microarray (PBM) and RNAcompete experiments. In chapter 4, I detail the contributions I made to the original affinity

regression paper, using affinity regression to model the interactions between the k-mer features of DNA/RNA probes and the K-mer amino acid features of transcription factors/RNA binding proteins that predict the binding affinities in Protein Binding Microarray (PBM) and RNAcompete experiments.

1.2.1 Statistical data analysis of tumor-stromal co-culture drug screen

I performed a statistical analysis of the co-culture data set from Straussman et al. [41] to establish feasibility of using affinity regression to model tumor-stromal interactions mediating resistance against anti-cancer therapies. In these co-culture experiments, co-culturing with stromal cells rescues cancer cells from drug-mediated killing. To determine which drugs are potential candidates for our study on innate drug resistance, I calculated a mean rescue score of the co-cultures treated with different drug dosages to determine which drug and dosage combination gives the best rescue of cancer cell lines by stromal cell lines. I looked at cancer proliferation in the PLX4720 treated co-cultures to get an idea of what drug resistance looks like in a positive and negative control cases– in melanoma where drug resistance is known to occur and in colorectal cancer where there is no drug resistance [1]. I looked at the relative proliferation of the cancer cells with stroma against the proliferation of cancer cells without stroma in melanoma and colorectal cell lines treated with different dosages of PLX4720.

Correlation analyses of the tumor-stromal co-culture drug screen

To identify of the cytokines correlated with strong drug resistance in a co-culture system, melanoma cell lines with BRAF V600E mutations treated with PLX4720, I repeated the statistical analysis performed by Straussman et al. [41], correlating the average rescue scores of melanoma cell lines treated with PLX4720 with individual cytokine expression levels secreted by the stromal cell lines. I confirmed that HGF is the cytokine most strongly correlated with resistance to PLX4720 in melanoma cell lines (Fig. 2.5,2.6)). To see if there are cytokines other than HGF that contribute to rescue in melanoma cell lines treated with PLX4720, I performed another analysis, correlating the individual rescue scores of the melanoma cell lines treated with PLX4720 with the cytokine expression levels secreted by the stromal cell lines (Fig. 2.7, 2.8, 2.9, 2.10). I found evidence that there may be other cytokines mediating resistance in melanoma cell lines treated with PLX4720. Finally I performed a correlation analysis across co-cultures treated with 15 different anti-cancer drugs. I calculated the correlation of the secretion level of the cytokines and the rescue scores for cancer cell lines and calculated an average correlation score for each subtype of cancer. I found cytokines for each cancer type and each drug dosage that were highly correlated with drug resistance.

1.2.2 Modeling tumor-stromal interactions that mediate innate resistance to cancer therapies

Both innate and acquired resistance to molecularly targeted therapies represent major challenges to cancer treatment. The importance of the tumor microen-

vironment in cancer initiation and progression is well established [26], but we are only beginning to understand the contributions of the microenvironment to therapeutic response and innate drug resistance. To study this issue computationally, I used affinity regression to model the effect of stromal cells on cancer cell drug sensitivity using a large published stromal-cancer co-culture data set consisting of 45 cancer cell lines, 23 stromal cells lines, and 35 drugs [30]. I represented each stromal cell by a feature vector of expression levels of secreted cytokines, measured by cytokine array in monoculture. I represented each cancer cell by pathway scores derived from curated signaling pathway databases, giving a view of the cellular circuitry that could receive and transduce signals from stromal cells. For each drug, the algorithm trained a regularized bilinear regression model that predicted the stromal rescue score for a cancer cell line from the pair of stromal and cancer feature vectors. By analysis of the trained model, I identified cytokines secreted by the stromal cell lines that may interact with signaling pathways in the cancer cells to mediate rescue. I found that affinity regression outperformed nearest neighbor approaches for the task of predicting rescue scores in cross-validation experiments. Further, for the BRAF inhibitor PLX4720, I confirmed that HGF is the cytokine most predictive of increased cancer cell proliferation in co-culture, and HGF participates in re-activating c-MET and PI3K/AKT signaling, consistent with published experimental reports [48]. Furthermore I found that HGF plays a similar role in afatinib and erlotinib treated NSCLC cell lines, while this interaction is not seen in NSCLC cell lines treated with the general cytotoxic drug docetaxel. The model of tumor-stromal interactions may lead to new insights into the role of stromal cells in promoting drug resistance and ultimately into how to treat cancer with combination therapies that target both the tumor and the microenvironment.

I compare multi-task affinity regression to an alternative method multi-task pairwise SVM. I describe a model that can be seen as the alternative to the multi-task affinity regression that uses classification rather than regression for modeling biological interaction data. I model the pairwise molecular interactions between the stromal cells and the tumor cells to explain innate drug resistance observed in the tumor-stromal co-culture drug screen using a multi-task pairwise SVM. I compare the performance of the two methods using 10-fold cross-validation trained on the tumor-stromal co-culture drug screen data treated with 15 anti-cancer therapies.

1.2.3 Application of affinity regression to PBM and RNAcompete data and comparison to multi-task pairwise SVM

In this chapter, I detail the contributions I made to the original affinity regression paper in 2015 [40]. There, affinity regression was used to model protein binding microarray (PBM) or RNAcompete experiments to learn family-level binding models for transcription factors (TFs) and RNA binding proteins (RBPs). There, we learned an interaction model between k-mer features of the nucleic acids and the K-mer features of the proteins. In this project, I prepared the K-mer feature matrices of the two inputs and the matrix of the binding affinities for the RNA binding protein analysis. I generated motifs for the affinity regression models for the Z-score affinity regression model for Z-scores obtained from PBM and RNAcompete experiments using the PWM-Align-Z algorithm. I plotted an example of the predicted and experimental Z-scores from the Z-score affinity regression model. I visualized the motifs for the RNA binding proteins

model in a circularized phylogenetic tree. I compared the motif accuracy of the motifs generated from affinity regression with the motifs generated from a nearest neighbor competitor algorithm by calculating the Kullback-Leibler divergence (D_{KL}) between the motifs generated from the models and the motifs calculated directly from the data, which we considered to be ground truth. Learning from PBM and RNAcompete data, the affinity regression model predicted the binding affinities of held-out proteins and identified key DNA/RNA-binding residues associated with binding. More generally, affinity regression can be used to model biological interaction data as the interactions between the features of two inputs, and it is possible to apply affinity regression to model and predict paired macromolecular or cellular interactions in any setting where there is a high-throughput affinity readout.

CHAPTER 2

STATISTICAL DATA ANALYSIS OF TUMOR-STROMAL CO-CULTURE DRUG SCREEN

2.1 Preliminary studies

Rescue scores of stromal-mediated drug resistance

In order to establish the feasibility of using affinity regression to model tumor-stromal interactions mediating drug resistance, we performed an analysis of the tumor-stromal co-culture drug screen data taken from Straussman et al. [41]. (Fig 2.1) shows an example of a co-culture experiment from the drug screen. For each cancer cell line the figure shows (HT-29), a number of stromal cell lines (HDF, HUVEC, LL86, Wi-38) which are grown in co-culture with and without the drug (Gemcitabine). There are quadruplicates for each condition and the measurements of GFP labeled cancer cell counts are averaged over these replicates.

The effect of stromal cells on cancer cells under drug conditions were first normalized by their effect on cancer cells without the drug. For example, from the labels in Fig 2.1 the effect of the HDF stromal cells on the HT-29 cancer cells is given by B/A, that is, the count of cancer cells with drug divided by count of the cancer cells without drug. Then the rescue score was calculated as the relative proliferation with stroma minus the relative proliferation without stroma. The rescue score of cancer cell line HT-29 by stromal cell line HDF is given below:

$$\text{Rescue(HDF confers to HT-29)} = \frac{B}{A} - \frac{D}{C}$$

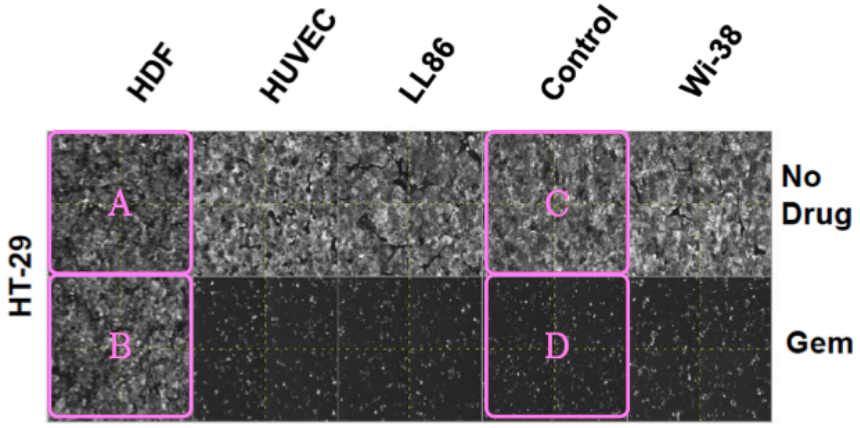

Figure 2.1: **Rescue of colorectal cancer cell lines from gemcitabine (Gem) by HDF stromal cells.** HT-29 was treated with 0.02 μ M Gemcitabine. Fluorescence microscopy looking at GFP positive cancer cells at day 7. All 4 quadruplicate wells are shown.

2.1.1 Mean drug rescue score in co-cultures treated with different drug dosages

We ran a preliminary statistical analysis of the co-culture drug screen. To identify drug and dosage combinations that exihibit stromal-mediated rescue of cancer cell lines. We calculated an average rescue score for each drug dosage, averaging across all the cancer and stromal co-cultures that were performed for that drug dosage (Fig. 2.2). We found that stromal cell lines mediate rescue of cancer cells in many targeted therapies but not for chemotherapeutic drugs. Chemotherapeutic drugs such as docetaxel, fluorouracil, caboplatin, paclitaxel, doxorubicin, gemcitabin, etoposize all have low mean rescue score while targeted therapies imatinib, erlotinib, afatinib, gefetinib, vemurafenib all have a higher range of rescue scores. We found the mean rescue score for PLX4720 in-

creases as the dosage of PLX4720 increases. Some of the other drugs of interest include SB590885, another BRAF inhibitor used in the treatment of melanoma. Since we have several drugs with high mean rescue scores that are either BRAF inhibitors or MEK inhibitors used in melanoma and we have validation data pertaining to PLX4720 from the original study, we decided to focus our initial studies on PLX4720 treated melanoma cell lines.

2.1.2 Dosage comparison of stromal mediated rescue in melanoma and colorectal co-cultures treated with PLX4720

We plotted the relative proliferation of the cancer cells with stromal cells against the proliferation of cancer cells without stromal cells in 7 melanoma cell lines treated with PLX4720 (Fig. 2.3). Here the distance from the diagonal indicates the magnitude of rescue or sensitization for the drug. We see that higher doses of the drug cause more killing of cancer cells and therefore lower cell counts increasing the potential for rescue stromal cells. There is also a bimodality in the relative proliferation for each set of co-cultures treated with the drug dosage. We see that for a subset of our cancers the drug has little effect, and there is also a cluster of cultures with low relative proliferation where the drug has a large effect. Stromal cells tend to mediate stronger rescue in the higher doses of PLX4720.

We also plotted the relative proliferation of cancer cells with stroma against the proliferation of cancer cells without stroma in 5 colorectal cancer cell lines treated with PLX4720 (Fig. 2.4). There is less drug efficacy of PLX4720, since

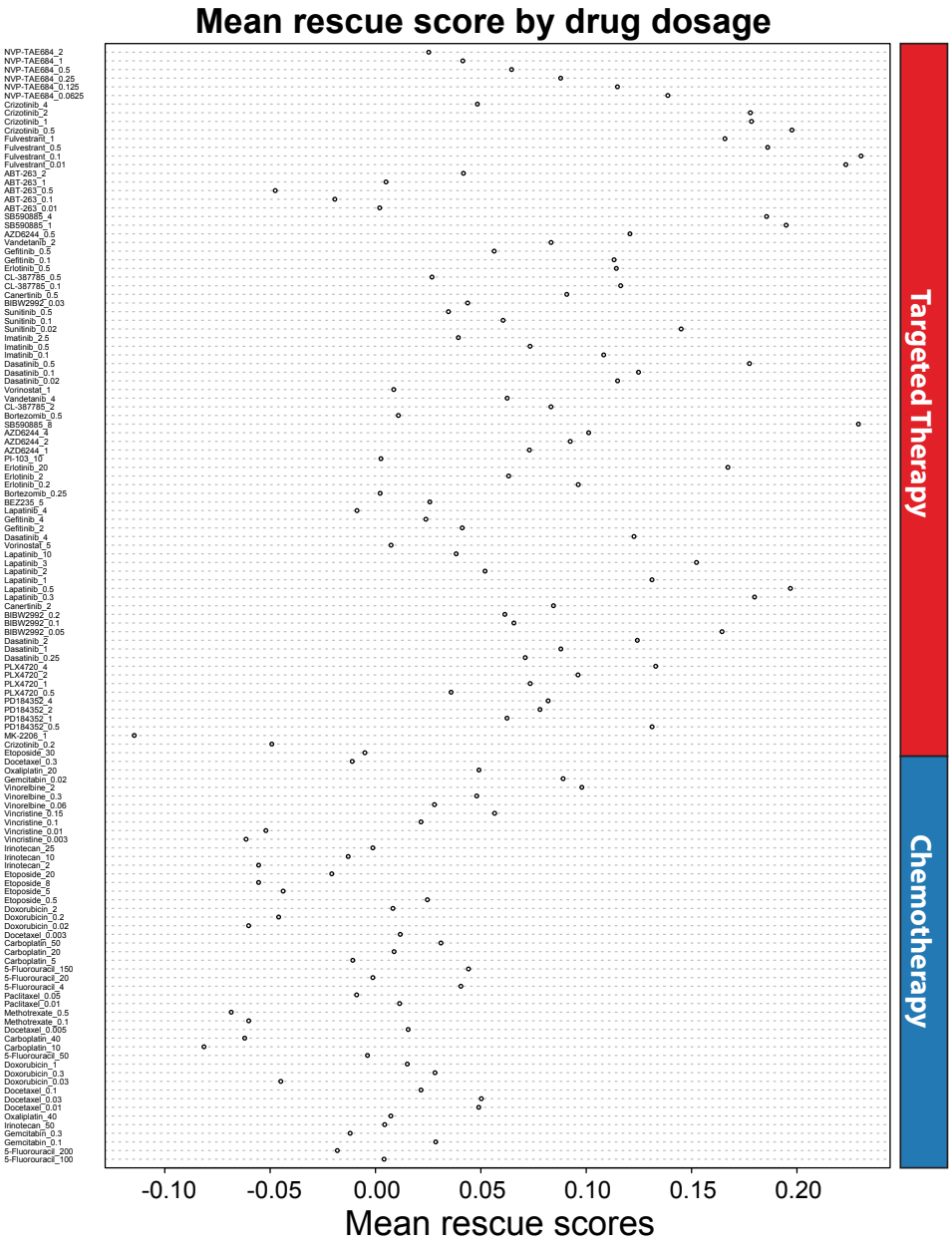

Figure 2.2: **Mean drug rescue score averaged across all tumor-stromal co-cultures indexed by the drug dosage.** The average rescue score for co-cultures treated was calculated and plotted for each drug dosage.

colorectal cell lines continue to proliferate even in the presence of PLX4720. We see that rescue is much greater across melanoma cell lines than across colorectal cell lines where even in cell lines sensitive to the drug with few examples far from the diagonal. This shows there is a difference in the rescue conferred by the stromal cell lines to the two different types of cancer cell lines treated with PLX4720, and stromal cells tend to elicit drug resistance in melanoma cells against PLX4720.

2.2 Correlation Analysis

2.2.1 Reanalysis of the correlation in cytokine arrays with average rescue score of melanoma cell lines treated with PLX4720

To identify the cytokines that are potentially eliciting innate resistance in cancer cells to targeted drug therapies, we performed a correlation analysis, correlating cancer cell line rescue scores with the cytokine expression levels secreted by stromal cell lines. We first calculated the correlations between proteins secreted by stromal cells and stromal average melanoma rescue scores R_M (the average rescue score over 7 melanoma cell, one for each stromal cell line), reproducing the results in Straussman et al. to make sure our data analysis is consistent. Then we performed the correlation analysis with individual cell lines, treating dosages for each drug separately, and computed the distribution of correlation

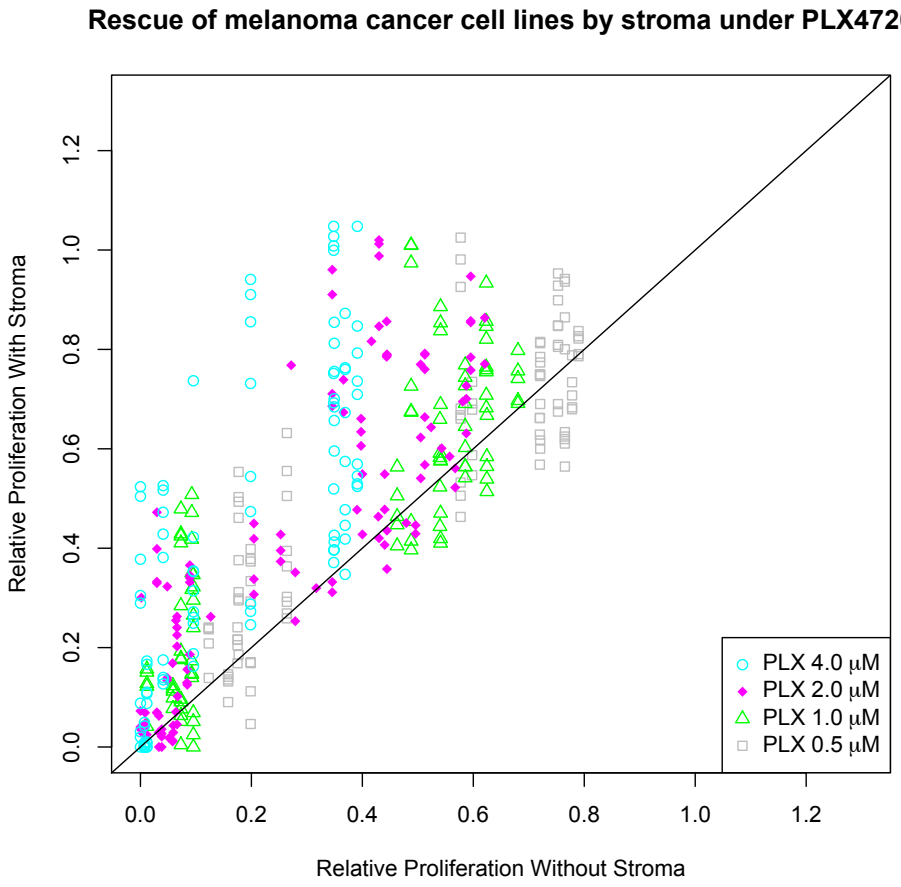

Figure 2.3: **Relative proliferation with stroma versus relative proliferation without stroma in melanoma cancer cell lines.** The relative proliferation of the cancer cells in co-culture with melanoma cell lines was plotted against the relative proliferation of cancer cells in monoculture for cancer cell lines treated with vemurafenib.

scores across the different dosages.

The stromal cell cytokine data set consisted of two types of antibody arrays: a Human Cytokine Array G4000 and a Biotin Label-based Human Antibody array were used to measure proteins secreted into the media by each of 18 stromal cell types [41]. The arrays measure 274 and 507 secreted proteins respectively.

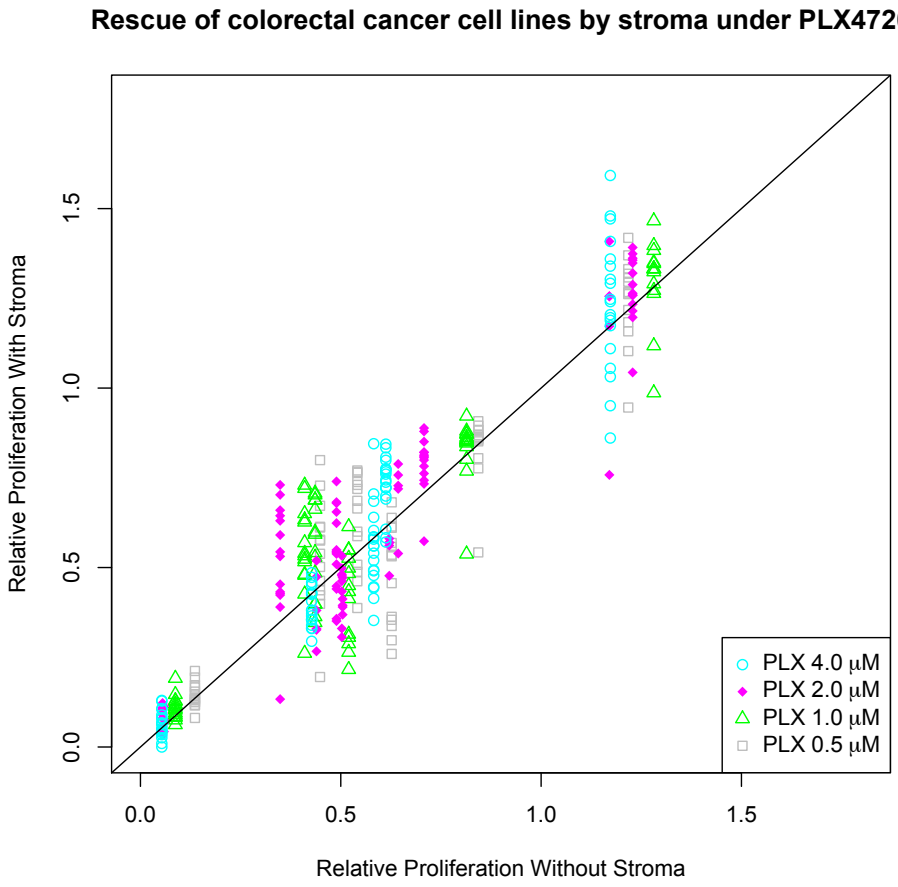

Figure 2.4: **Relative Proliferation with stroma versus relative proliferation without stroma in colorectal cancer cell lines.** The relative proliferation of cancer cells in co-culture with colorectal cell lines was plotted against the relative proliferation of cancer cells in monoculture for cancer cell lines treated with vemurafenib.

The resulting antibody arrays were normalized using internal controls and the values of cytokines in clear media containing 10% FBS were subtracted. After filtering out the low level cytokine expression values (expression of >50 in at least one stromal cell line), measurements for 85 and 414 cytokines remained from each array. A correlation coefficient was calculated for each cytokine between its log2-transformed secretion level by all 18 stromal cell lines and the

average melanoma rescue score of each of these cell lines. The mean melanoma rescue effect of each stromal cell line (R_M) was calculated by averaging the rescue scores of the this cell line across all melanoma cell lines and all PLX4720 concentrations. If a drug did not reduce proliferation to below 0.3 for a cancer cell line in monoculture, we removed this drug/cell line pair from the analysis. After this filtering step, for each cytokine, we correlated the melanoma rescue score over 18 stromal lines (R_M) with the expression level of the cytokines across these stromal lines. HGF is a prominent cytokine with high correlation with the mean stromal cell line's rescue score, but there were also other cytokines with high correlations such as GDF9 and uPA (with positive correlation) and IL-3 (with negative correlation) (Fig. 2.5, 2.6).

2.2.2 Analysis of the correlation in cytokine arrays with rescue scores in individual melanoma cell lines treated with PLX4720

To examine if there are cytokines other than HGF that contribute to rescue, we proceeded to perform the correlation analysis with the individual melanoma cell lines, instead of the averaged stromal rescue score. We also treated the drug dosages separately and looked at the distribution of correlation scores across different dosages (Fig. 2.7, 2.8, 2.9, 2.10). Here we show the correlation of HGF cytokine expression level with PLX4720 at $1\mu M$ and 2 μM. We see that HGF under the drug PLX4720 does score a high median correlation in the cancer cell lines but is not the only cytokine correlating with the rescue scores. Also there are cytokines with high negative correlation with rescue such as IL-6 and IL-3

, suggesting that when these cytokines are present, the targeted therapy may work more effectively. Other cytokines such as GDF9 and VEGF (with positive correlation) are potentially interesting as well.

2.2.3 Correlation analysis for cancer subtypes across co-cultures treated with anticancer drugs

We performed a correlation analysis across 15 anti-cancer drugs and 44 cytokines in the smaller cytokine array. 44 cytokine measurements were obtained from the normalized cytokine array filtered with an expression cut off of 0.75 after normalization. We calculated the correlation of the stromal cytokines and the rescue scores for cancer cell lines across the stromal cell lines. For each cancer type, we averaged the correlations across all the cancer cell lines of that type to create an average correlation score. For 44 cyokines in the normalized cyokine array and across all active drug dosages, we produced a ranked list of pearson correlation scores for different types of cancer. From our ranked list, the highest scoring cytokine was GM-CSF whose cytokine secretion was highly correlated with rescue in melanoma in co-cultures treated with SB590885 $4\mu M$ with a correlation 0.70 (Fig. 2.11). The next two highest scoring cytokines were HGF in melanoma cell lines treated with PLX4720 4 μM and 2 μM whose correlations were both 0.68 (Fig. 2.12, 2.13). The fourth highest scoring cytokine was MMP-10 in melanoma cell lines treated with PLX4720 1 μM with correlation 0.65 (Fig. 2.14). In breast cancer cell lines, the highest scoring cytokine was VEGF-C in co-cultures treated with afatinib 0.05 μM (r= 0.58) (Fig. 2.15). In head and neck cancer cell lines the highest scoring cytokine was MMP-10 in lapatinib 4 μM

treated cell lines (r= 0.57) (Fig. 2.16). In non-small cell lung cancer the highest scoring cytokine was HGF in afatinib 0.03 μM treated cell lines (r = 0.54) (Fig. 2.17).

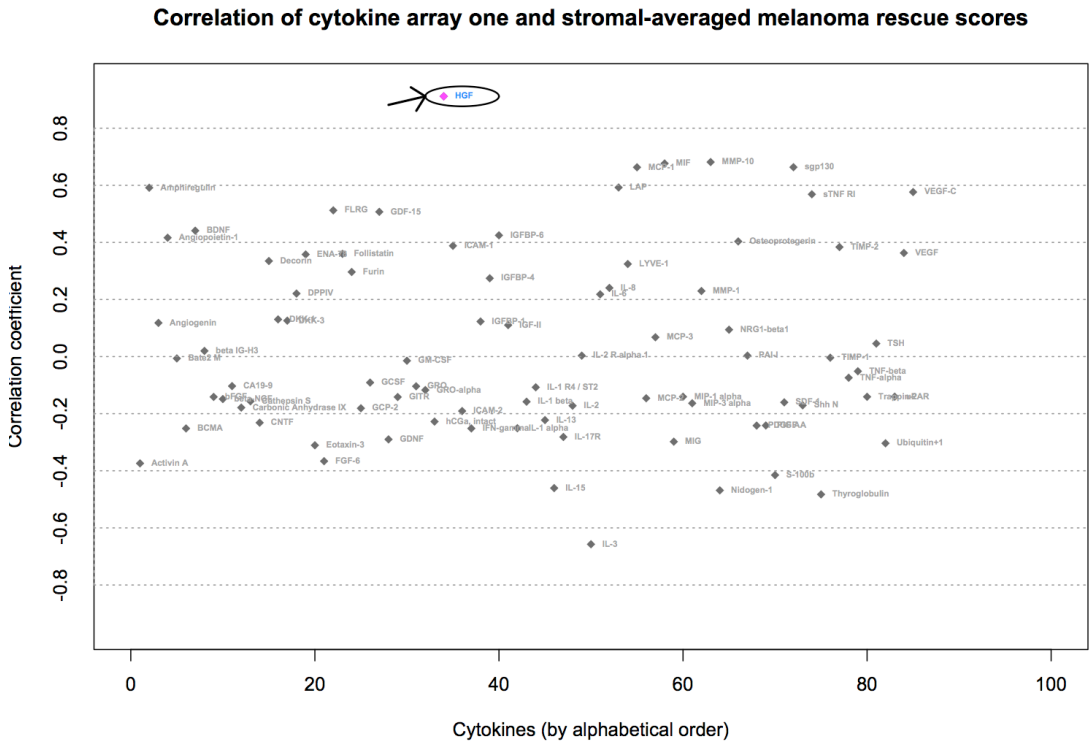

Figure 2.5: **Correlations for each of the cytokines in cytokine array one with stromal average melanoma rescue score R_M.** The correlation of the secretion of each cytokine in cytokine array one with the stromal average rescue score for melanoma cell lines treated with PLX4720 was calculated. The stromal average rescue score for melanoma was calculated for each of the stromal cell lines as the average rescue score across all melanoma cell lines in co-culture with the stromal cell line.

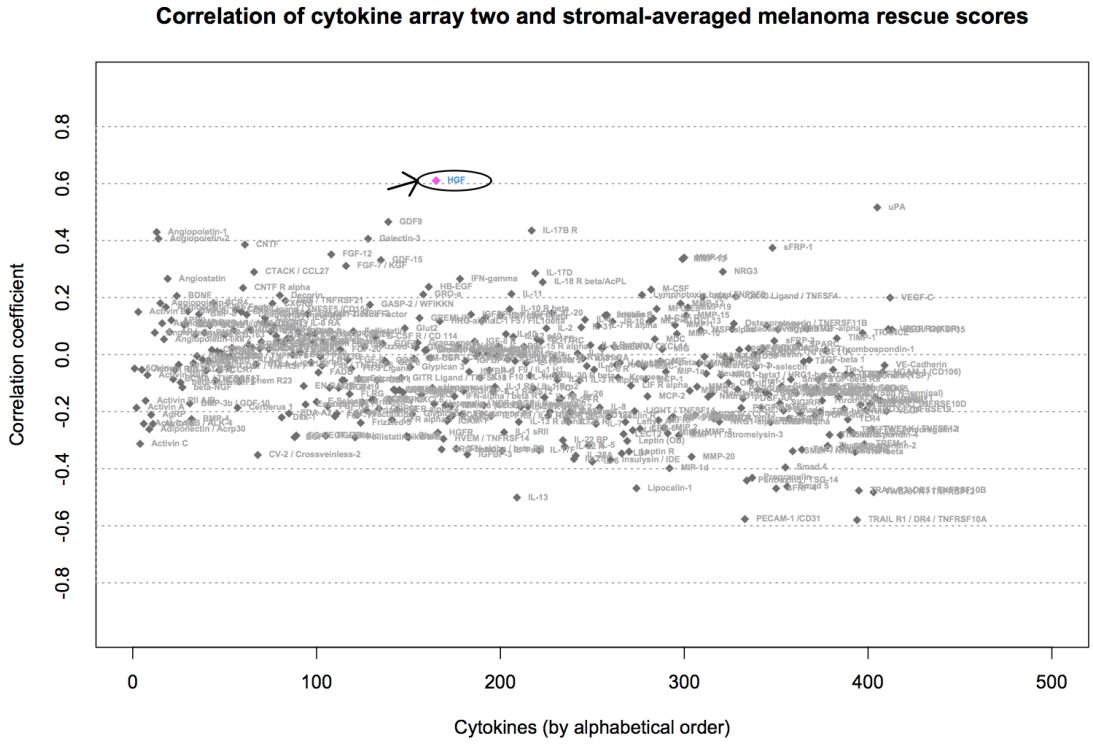

Figure 2.6: **Correlations for each of the cytokines in cytokine array two with stromal average melanoma rescue score R_M.** The correlation of the secretion of each cytokine in cytokine array one with the stromal average rescue score for melanoma cell lines treated with PLX4720 was calculated. The stromal average rescue score for melanoma was calculated for each of the stromal cell lines as the average rescue score across all melanoma cell lines in co-culture with the stromal cell line.

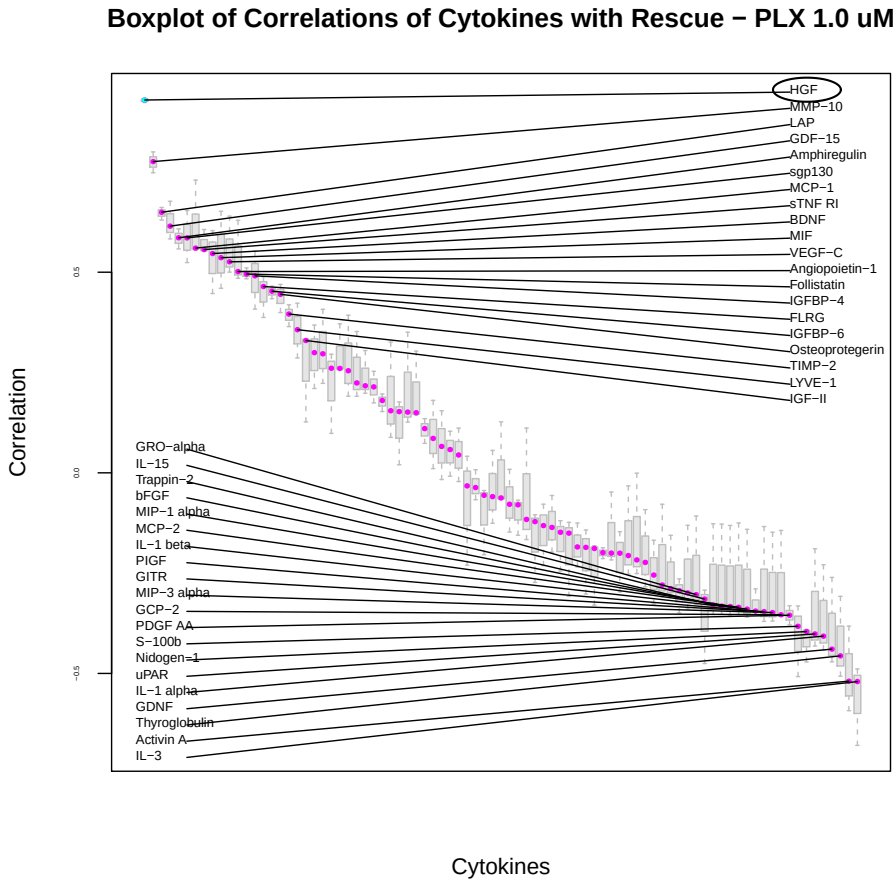

Figure 2.7: **Box plots of correlations of cytokines secretion in cytokine array one with rescue of melanoma cell lines treated with PLX4720 1 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array one with the rescue scores from the melanoma cell lines under the drug PLX4720 1 μM.

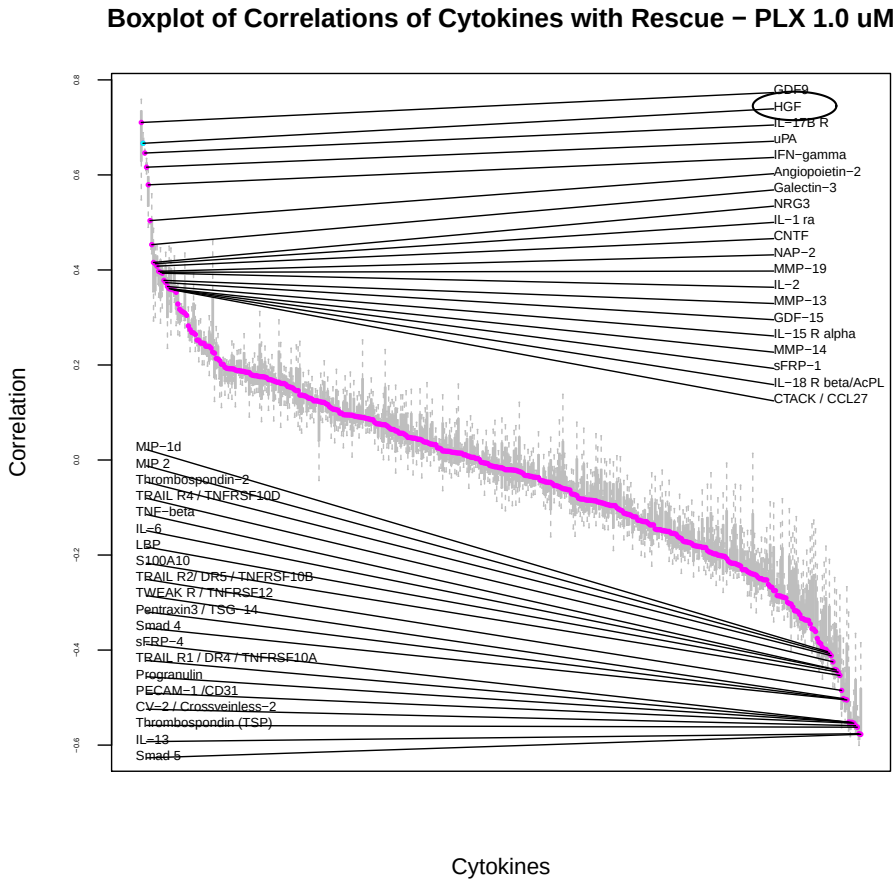

Figure 2.8: **Box plots of correlations of cytokines secretion in cytokine array two with rescue of melanoma cell lines treated with PLX4720 1 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array two with the rescue scores from the melanoma cell lines under the drug PLX4720 1 μM.

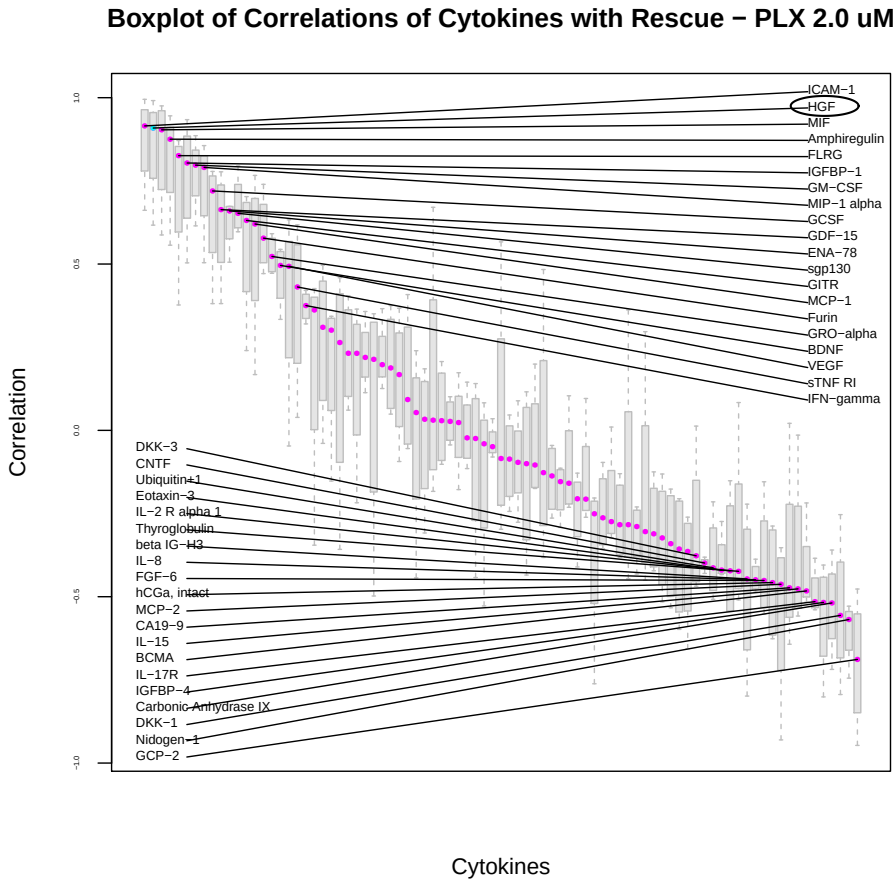

Figure 2.9: **Box plots of correlations of cytokines secretion in cytokine array one with rescue of melanoma cell lines treated with PLX4720 2 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array one with the rescue scores from the melanoma cell lines under the drug PLX4720 2 μM.

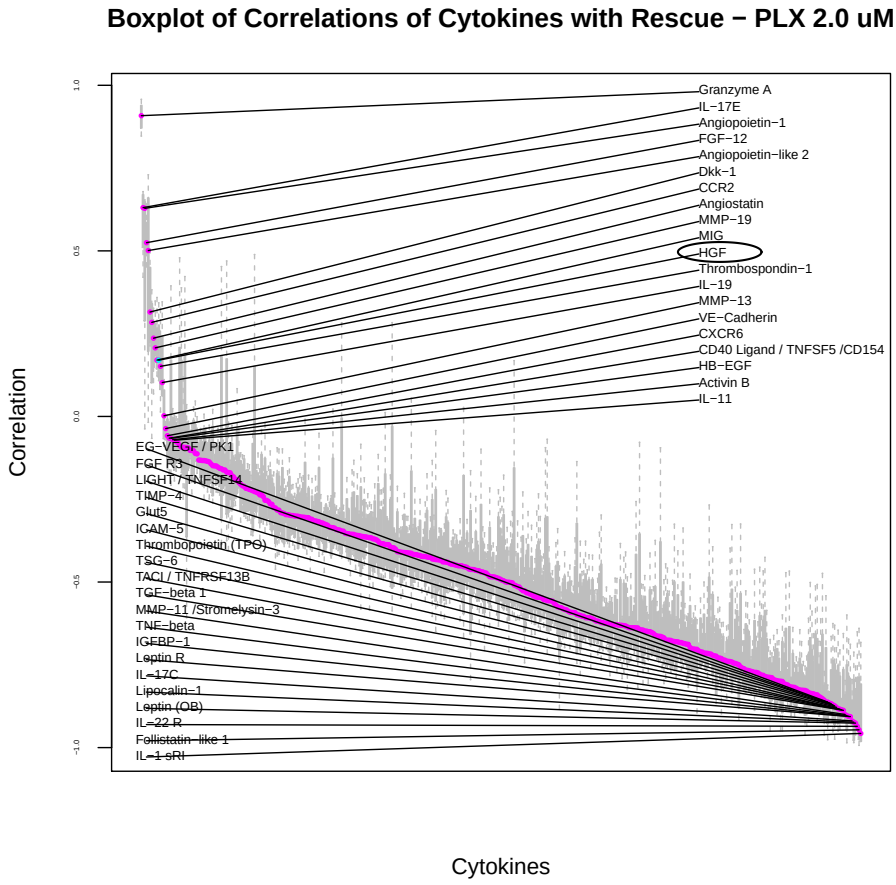

Figure 2.10: **Box plots of correlations of cytokines secretion in cytokine array two with rescue of melanoma cell lines treated with PLX4720 2 μ M.** Here we plot box plots of the correlations of each of the cytokines in cytokine array two with the rescue scores from the melanoma cell lines under the drug PLX4720 2 μM.

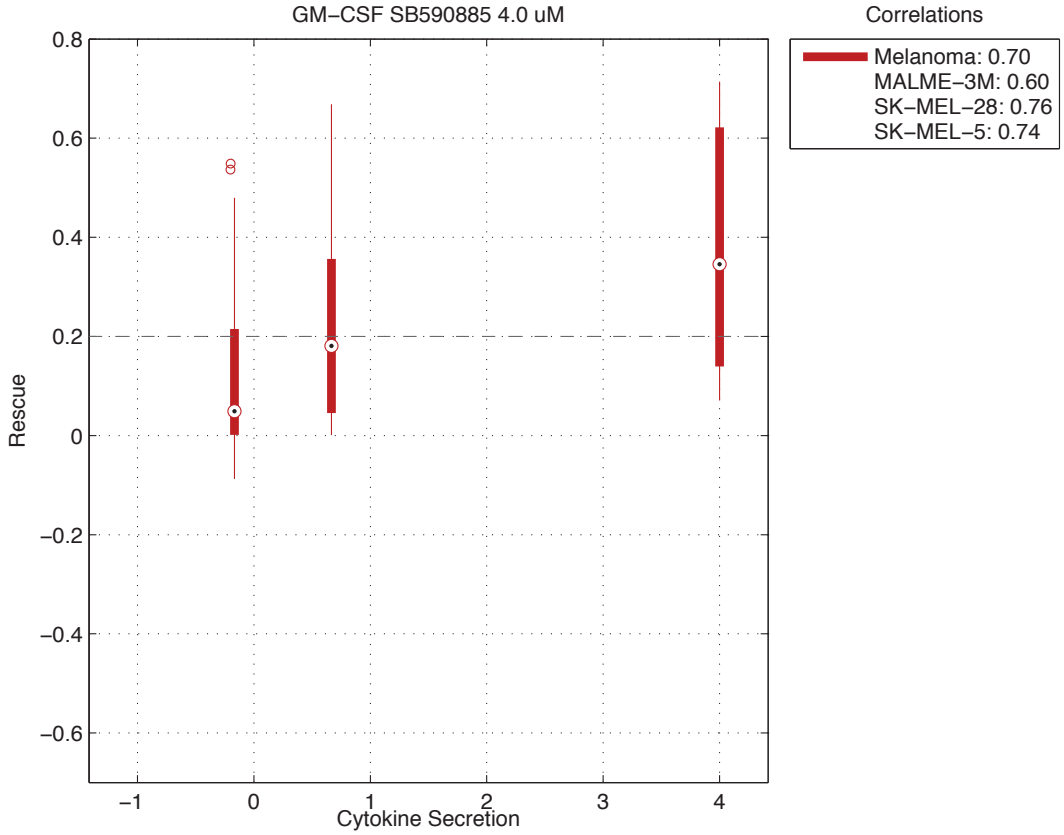

Figure 2.11: **Box plots of rescue scores plotted against GM-CSF secretion in SB590885 4 μM treated melanoma** The rescue scores are plotted as a function of the secretion of GM-CSF secretion in melanoma cell lines treated with SB590885 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines for the cancer subtypes is calculated.

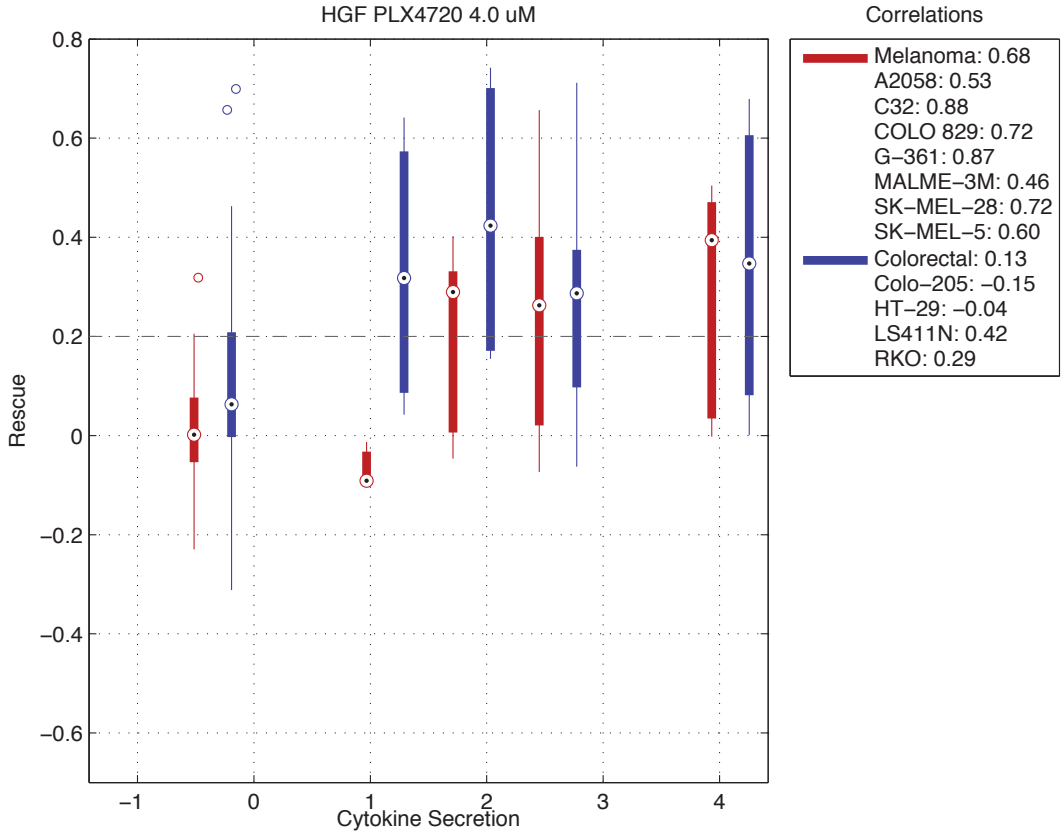

Figure 2.12: **Box plots of rescue scores plotted against HGF secretion in vemurafenib 4 μM treated melanoma** The rescue scores are plotted as a function of the secretion of HGF secretion in melanoma cell lines treated with PLX4720 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

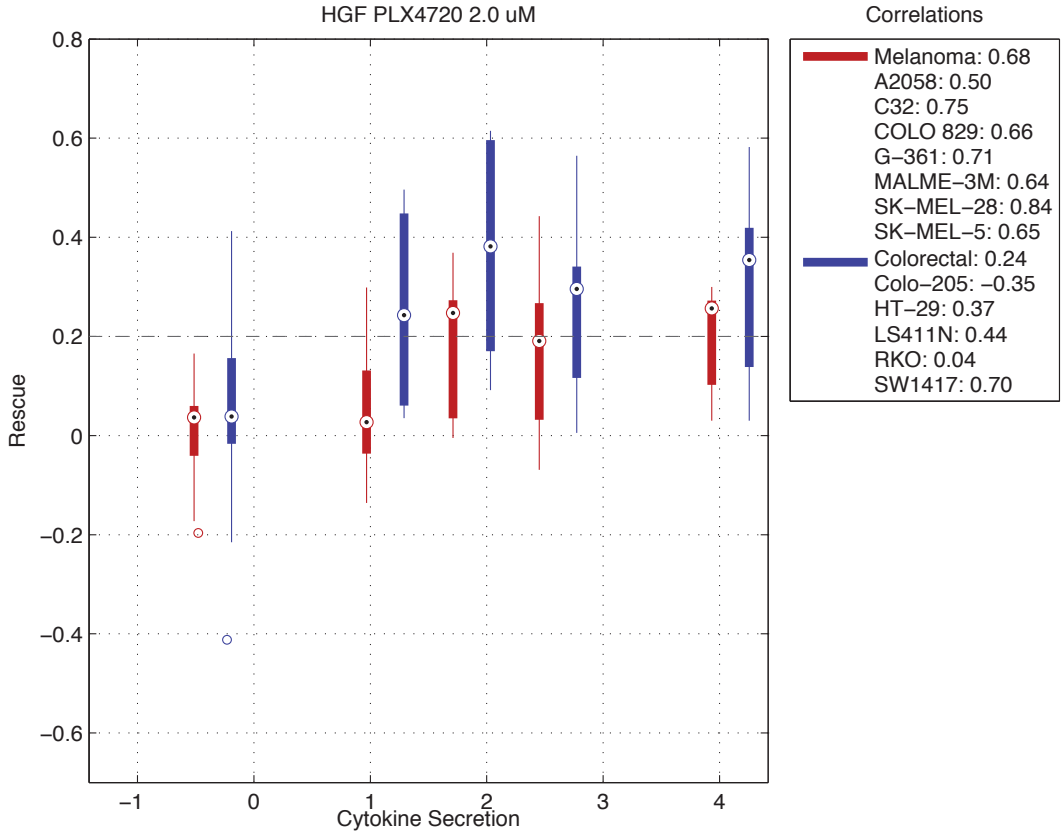

Figure 2.13: **Box plots of rescue scores plotted against HGF secretion in vemurafenib 2 μM treated melanoma** The rescue scores are plotted as a function of the secretion of HGF secretion in melanoma cell lines treated with PLX4720 2 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

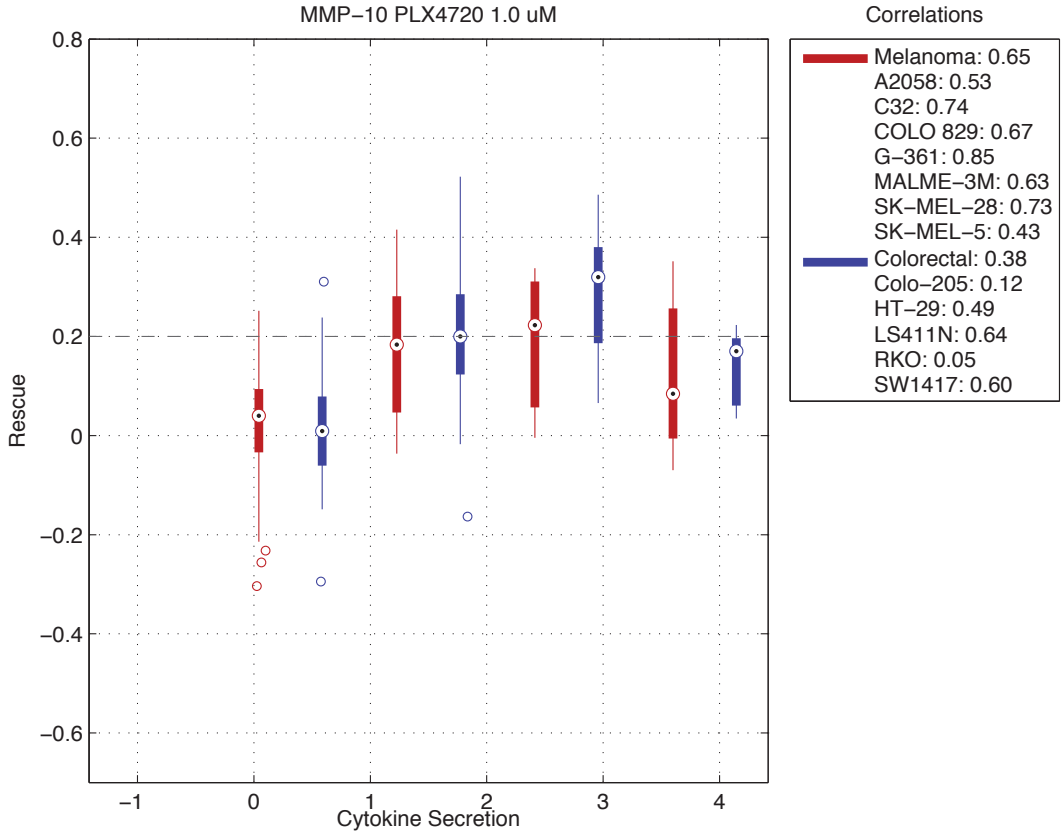

Figure 2.14: **Box plots of rescue scores plotted against MMP-10 secretion in vemurafenib 1 μM treated melanoma** The rescue scores are plotted as a function of the secretion of MMP-10 secretion in melanoma cell lines treated with PLX4720 1 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

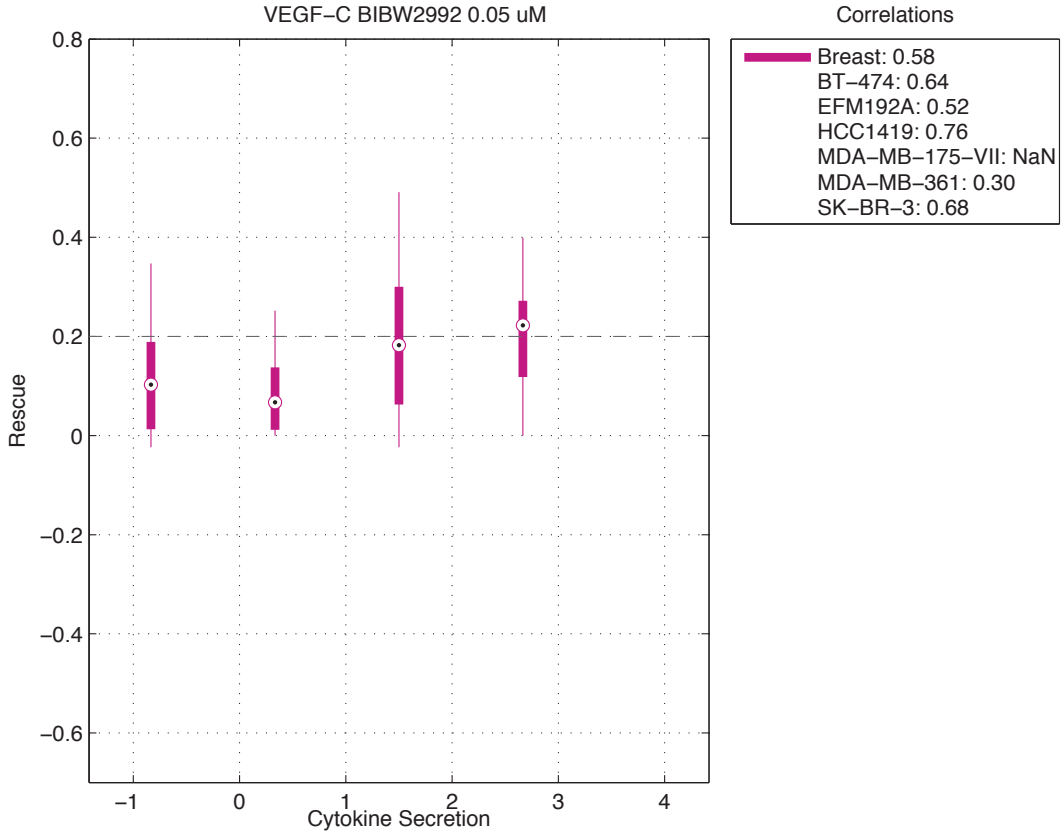

Figure 2.15: **Box plots of rescue scores plotted against VEGF-C secretion in afatinib 0.05 μM treated breast cancer** The rescue scores are plotted as a function of the secretion of VEGF-C secretion in breast cancer cell lines treated with afatinib 0.05 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

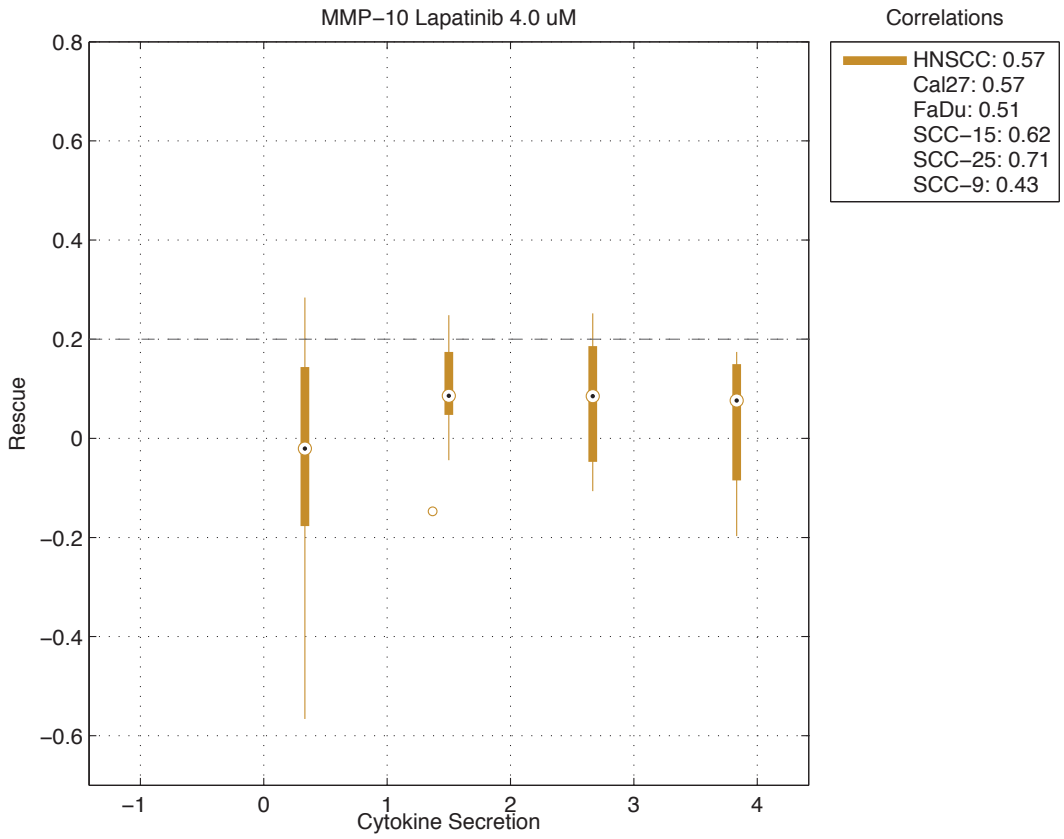

Figure 2.16: **Box plots of rescue scores plotted against MMP-10 secretion in lapatinib 4 μM treated head and neck cancer** The rescue scores are plotted as a function of the secretion of MMP-10 secretion in head and neck cancer cell lines treated with lapatinib 4 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

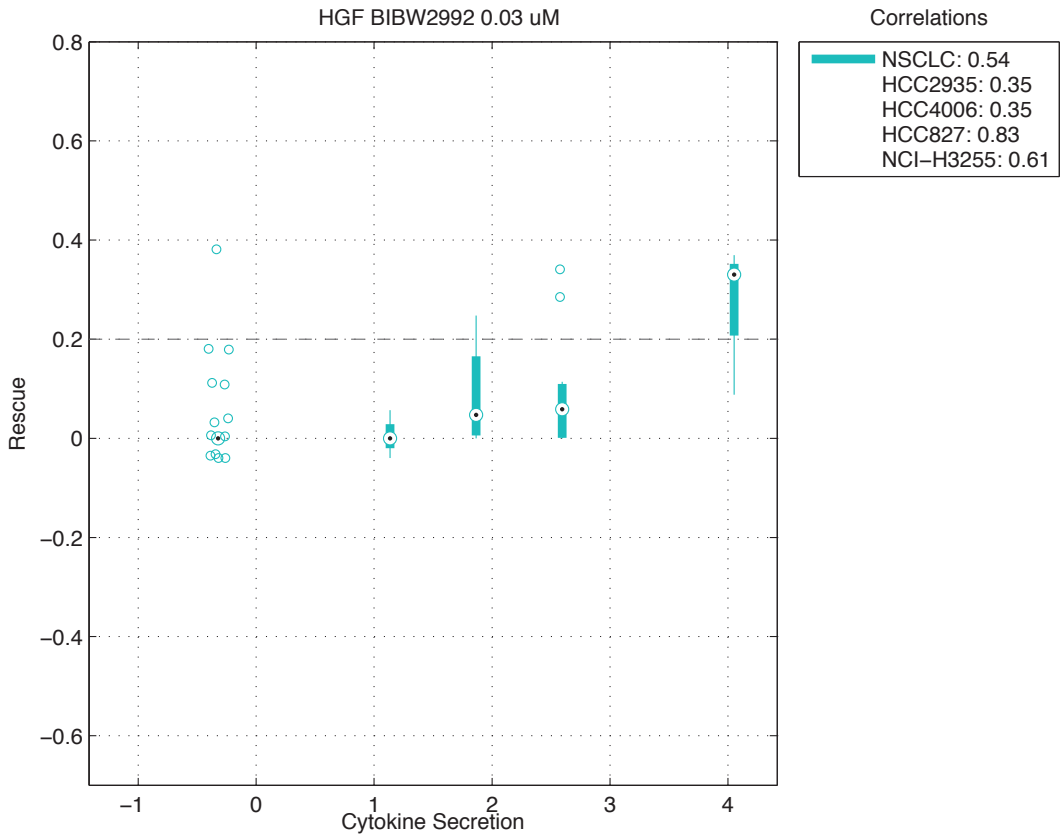

Figure 2.17: **Box plots of rescue scores plotted against HGF secretion in afatinib 0.03 μM treated non-small cell lung cancer** The rescue scores are plotted as a function of the secretion of HGF secretion in non-small cell lung cancer cell lines treated with afatinib 0.03 μM. The average correlation between the secretion of the cytokine and rescue scores for cancer cell lines and for the cancer subtypes is shown.

CHAPTER 3

MODELING TUMOR-STROMAL INTERACTIONS THAT MEDIATE INNATE RESISTANCE TO TARGETED CANCER THERAPIES

Portions of this chapter first appeared in Yang et al. [30] and were written in collaboration with Christina Leslie. [1]

3.1 Introduction

3.1.1 Background

Cancer is a class of proliferative diseases thought to arise due to a series of somatic mutations that result in the dysregulation of cellular pathways. In addition to cell-intrinsic factors, cancers take place within the context of and are known to be influenced by extrinsic factors of the surrounding host microenvironment. The microenvironment, comprised of innate and adaptive immune cells, macrophages, fibroblasts, the blood and lymphatic vascular networks and the extracellular matrix, has been implicated in multiple stages of cancer development such as cancer initiation, progression, and metastasis. Cancer cells have been known to recruit stromal cells in the surrounding microenvironment to participate in interactions such as reciprocal paracrine signaling that promote cancer growth and metastatic dissemination to distant organs [26].

Targeted drug therapies, in which a specific molecular target within the context of a cellular mechanism is blocked by a small molecule, represent a sig-

[1] As per the Cornell dissertation guidelines, the dissertation can include material that has been previously published or is soon to be published.

nificant advance in personalized medicine in the genomics era. Oncoprotein-targeted drug therapy offers a potentially very promising option for cancer treatment. However, resistance to targeted drug therapies poses a major challenge to the future success of these treatments. Clinical trials of two recently popular drugs: imatinib (targeting cells with KIT mutations) and vemurafenib (targeting BRAFV 600E mutations) both showed patients had marked response to the drug: 50% response in the case of vemurafenib [22, 27]. These responses, however, were partial and in most of the cases resistance emerged after a short period of time eventually leading to cancer relapse. In the case of vemurafenib tumors recurred within six months of treatment. These short term relapses suggest that mechanisms exist to render a substantial portion of tumor cells resistance to treatment.

Two recent studies have implicated that innate drug resistance may, in part, be caused by the factors secreted by the tumor microenvironment [41, 48]. The study by the group at the Broad Institute screened capacity of 23 stromal cell lines to alter the response of 45 cancer cell lines to 35 commonly used anti-cancer agents. Of the 23 targeted agents in the panel, there was evidence of microenvironment-mediated resistance to 15 (65%). The study by the group at Genentech examined the effect on drug response to eight anti-cancer drugs of 41 cancer cell lines to exposure to six RTK ligands known to be widely expressed in tumors. The study showed HGF, FGF, NRG1, and EGF all displayed a protective effect on the cancer cell lines against the anti-cancer agent, whereas IGF and PDGF had relatively little effect. The two studies identified HGF as an important soluble factor that was capable of mediating resistance to BRAF and HER2 inhibitors. These findings suggest an important role for the tumor microenvironment in mediating innate drug resistance to molecularly targeted

therapies.

Although mechanisms of resistance to molecularly targeted therapies are still not fully understood, one model suggests that cancer cells that are subject to oncogene addiction can bypass the targeted mechanism through the activation of alternative pathways, reactivating oncogenic signaling [4, 35, 36, 37, 44, 48]. As cancer cells recruit stromal cells in the tumor microenvironment to promote cancer growth, stromal cells can potentially participate in this process of conferring resistance by secreting cytokines that can act as upstream activators of alternative pathways that can reactivate the oncogenic signaling inside the cancer cell [30].

Here we model and dissect the interactions between the tumor and stromal cells that mediate innate resistance to target therapies. We will attempt to understand how cytokines secreted by the stromal cells interact with and activate alternative pathways in the cancer cells to mediate innate drug resistance to molecularly targeted therapies. By developing a statistical model to predict how tumor-stromal interactions give rise to cancer cell proliferation, we will gain new insights into the role of stromal cells in promoting drug resistance and ultimately into how we might treat cancer with combination therapies that target the tumor microenvironment.

3.1.2 Prior Work

Both innate and acquired resistance to molecularly targeted therapies represent major challenges to cancer treatment. Recent studies have proposed a role for the tumor microenvironment in therapeutic response, demonstrating that cy-

tokines secreted by stromal cells can rescue cancer cells from killing by targeted drugs. To systematically study the stromal contribution to innate drug resistance, we used a method called affinity regression to model the effect of stromal cells on cancer cell drug sensitivity using a large published stromal-cancer coculture data set. Our model represents each stromal cell by the feature vector of expression levels of its secreted cytokines, and each cancer cell by pathway scores derived from curated signaling pathway databases, giving a view of the cellular circuitry that could receive and transduce signals from stromal cells. For each drug, our algorithm trains a regularized bilinear regression model that predicts the stromal rescue score for a cancer cell line from stromal and cancer cell features. We confirmed that affinity regression outperformed nearest neighbor approaches for the task of predicting rescue scores in cross-validation experiments. Furthermore, by analysis of the trained model, we identified cytokines secreted by stromal cells that may interact with signaling pathways in cancer cells to mediate rescue. For the BRAF inhibitor PLX4720, our model identified HGF as the cytokine most predictive of melanoma cancer cell rescue and associated with c-MET and PI3K signaling, consistent with published experimental reports. Our model also predicted that HGF plays a similar role in non-small cell lung carcinoma (NSCLC) cells treated with EGFR inhibitors, and we confirmed this prediction experimentally for afatinib and erlotinib. Our statistical model of tumor-stromal interactions may lead to new insights into the role of stromal cells in promoting drug resistance and could ultimately suggest combination therapies to target the tumor microenvironment.

Tumors are complex tissues that comprise cancer cells as well as diverse stromal cells of the tumor microenvironment (TME) including fibroblasts, endothelial cells, and immune cells along with the extracellular matrix they produce

[8]. Numerous studies over the past decade have established that the TME contributes to regulation of tumor initiation, progression, and metastasis [8], and recent work has begun to elucidate its role in modulating response to therapy [16]. TME-mediated drug resistance includes both innate resistance, resulting from a preexisting network of tumor-stromal interactions that promote survival/proliferation of cancer cells, and acquired TME resistance, where the therapeutic intervention leads to changes in the composition or state of cells of the TME that ultimately protect cancer cells.

In the past several years, large-scale projects in precision oncology have generated drug dose response data across large panels of molecularly characterized cancer cell lines, with the goal of training predictive models of anticancer drug response [24, 49]. However, these monoculture experiments do not model the stromal contribution to response to therapy, and therefore the predictive signatures trained on these data sets may have limited ability to generalize to patients. Recently, several studies have measured response to cancer therapeutics in cancer-stromal cell co-cultures to begin to address the issue of stromal-mediated innate resistance [41, 42]. While co-culture experiments cannot model the complex niche established by the TME, they do enable assessment of whether tumor-stromal paracrine signaling alters drug response. In particular, Straussman et al. generated and analyzed drug dose response data from cancer cell-stromal cell co-cultures treated with various cytotoxic and molecularly-targeted drugs and established that hepatocyte growth factor (HGF), a cytokine secreted by some of the stromal cell lines, could rescue BRAFV600E mutant melanoma cell lines from killing by the BRAF inhibitor PLX4720 (an analogue of which, vemurafenib, was recently approved by the US Food and Drug Administration (FDA) for the treatment of BRAF-mutant

melanoma) [41]. Another study by Genentech, published at the same time, came to similar conclusions [48].

Here we revisit the Straussman et al. data set [41] and ask if a more systematic analysis using machine learning modeling could identify additional secreted cytokine-drug pairs that mediate innate drug resistance in cancer cells. To do this, we applied a multi-task version of affinity regression, a supervised learning algorithm we recently developed for modeling pairwise biological interactions [21, 40], to drug co-culture experiments. Representing each stromal cell by a feature vector of its secreted cytokine expression levels and each cancer cell by a vector of mRNA expression derived pathway scores, we trained a regularized bilinear regression model for each drug to predict the rescue score for each cell line pair (stromal, cancer). Here, the rescue score (introduced previously, [41]) is a quantitative measure of the extent to which co-culturing with stromal cells enables increased cancer cell survival/proliferation in the presence of drug. We used multi-tasking to jointly train across different dosages of the same drug or different drugs in the same class.

Our affinity regression analysis recovered the finding that HGF can elicit innate resistance in melanoma cell lines treated with BRAF inhibitors. Additionally, we predicted that HGF could also mediate innate resistance in lung cancer cell lines treated with EGFR inhibitors, and we experimentally confirmed these predictions for afatinib and erlotinib, two EGFR inhibitors in clinical use in lung cancer. Our methodology provides a new strategy for mining drug sensitivity in co-culture data sets and for deciphering innate TME-mediated mechanisms of drug resistance.

3.2 Results

3.2.1 Multi-task affinity regression models tumor-stromal interactions in co-culture drug treatment experiments

To uncover stroma-mediated innate drug resistance mechanisms, we interrogated co-culture experiments from Straussman et al. [41]. To infer both the cytokines secreted by stromal cells and the signaling pathways in cancer cells that together mediate drug resistance, we used a supervised machine learning approach called affinity regression [40]. For each drug and tested dosage, we computed a rescue score, slightly modified from the original study (see Methods), for each stromal-cancer cell line combination. This score quantifies the proliferation advantage acquired by cancer cells when co-cultured with the stromal cells under drug treatment, compared to cancer cells grown in monoculture under drug treatment. We represented each stromal cell line by the feature vector of its secreted cytokine expression levels. Expression levels for different cytokines were not highly correlated. This suggests that it would be unlikely for multiple cytokines with a similar expression pattern to elicit the drug resistance observed in the co-cultures. For the cancer cell line representation, we used microarray expression data for the entire Cancer Cell Line Encyclopedia (CCLE) [24] to define scores for 75 pathways from the Pathway Interaction Database [6]. Pathways were also not highly correlated with each other.

Hierarchical clustering of all 1036 cell lines in CCLE by 75 pathway scores confirmed that pathway information can group cancer cells by their cell type of origin (Fig. 3.1). Individual pathway scores also provide meaningful char-

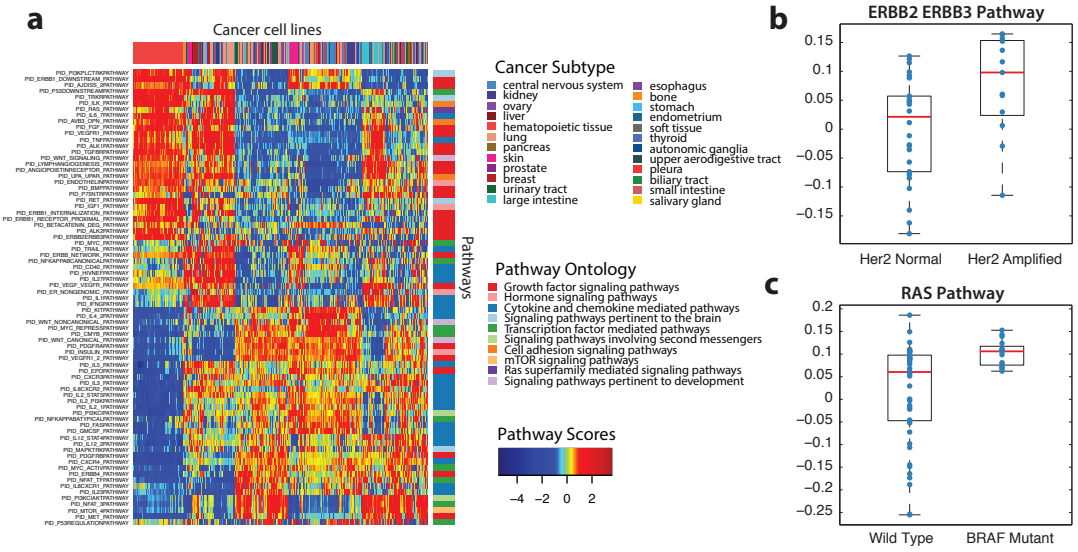

Figure 3.1: **Hierarchical clustering of the pathway scores groups cancer cell lines by tissue of origin** (a) Hierarchical clustering of the pathway scores for 75 curated pathways across 1036 cell lines from CCLE groups cancer cell lines by tissue of origin and clusters together growth factor signaling pathways (red) and cytokine/chemokine mediated pathways (blue). Hematopoietic, stomach, and large intestine cancer cell lines have high growth factor signaling pathway scores, while lung and skin cell lines have high cytokine and chemokine mediated pathway scores. (b) Her2 amplified breast cancer cell lines have higher ERBB2 signaling pathway scores than Her2 wild type breast cancer cell lines. (c) Melanoma cell lines with BRAF mutations have higher RAS/RAF signaling pathway scores than those with wildtype BRAF.

acterizations of cell line signaling states. For example, Her2-amplified breast cancer cell lines have higher ERBB2 signaling pathway scores than Her2 wild-type breast cancer cell lines (Fig. 3.1), while melanoma cell lines with BRAF mutations have higher RAS/RAF signaling pathway scores the BRAF wild-type melanoma cell lines (Fig. 3.1). The positive direction of the pathway score determined by principal component analysis (PCA) is given by the direction in which the majority of the genes are positive. In some cases the majority of the genes may not give the direction of the up-regulation of the pathway and in this case our sign convention would not be appropriate. Given this case, choice of the sign of the pathway is somewhat arbitrary.

We then used affinity regression to learn a bilinear regression model that explains the rescue scores as interactions between pathway scores of the cancer cells and cytokine features of the stromal cells (Fig. 3.2). Intuitively, the algorithm learns a weighting over interactions between cancer cell pathway features and stromal cell cytokine features that explains how these pathway-cytokine combinations contribute to the observed rescue data. Formally, we set up a bilinear regression problem to learn an interaction matrix W between cancer cells, represented by the input matrix C, and stromal cells, represented by the input matrix S, that reconstructs the output matrix Y of observed rescue scores (3.2). Each cancer cell line is represented by its pathway features as a row in C, and each stromal cell line by its cytokine features as a row in S; columns in Y represent the rescue scores of different cancer cells in co-culture with stromal cells. The affinity regression interaction model is formulated as:

$$CWS^T \approx Y \tag{3.1}$$

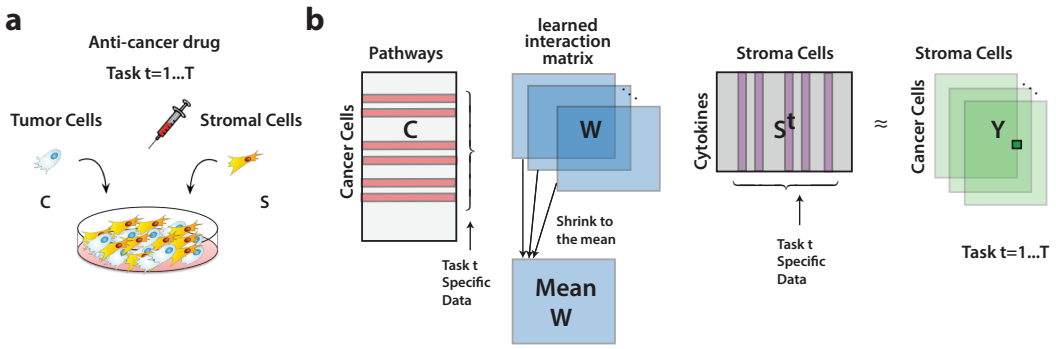

Figure 3.2: **Affinity regression predicts stromal-mediated rescue from targeted therapies from cancer cell line pathway scores and stromal cell cytokine data.** (a) Drug co-culture experiments quantify the extent of stromal-mediated rescue of cancer cells treated with targeted agents, assigning a rescue score to each (cancer cell line, stromal cell line, drug/dosage) combination. (b) Multi-task affinity regression learns to predict rescue scores from pathway score features of cancer cells and secreted cytokine levels of stromal cells using a regularized bilinear regression strategy. Each task trains on the co-culture experiments for a specific drug and dosage. The model is represented as an interaction matrix W_t between pathways and cytokines; S and C represent the feature matrices of cytokine expression values for stromal cells and pathway scores for cancer cells, respectively. Tasks corresponding to different dosages of the same drug or to drugs in the same class are jointly trained by shrinking model matrices W_t for different tasks towards the average task W_o.

where C, S, Y are known and W is unknown. We then convert the problem from a bilinear to a regular regression by taking a tensor product of the input matrices and solve for W with ridge regression. We perform multi-task affinity regression for 13 drugs with a focus on two specific drug groups: EGFR and BRAF/MEK inhibitors. Here, the multiple tasks consist of different dosages for a specific drug (most targeted drugs were tested at 4 different dose levels) or

different drugs within a class (e.g. multiple EGFR inhibitors). For each multi-task regression we add a constraint term to the optimization that shrinks the W_t task matrices towards a mean W_o matrix.

Figure 3.3: **Cytokine and pathway mappings** (a) Multiplying a cancer cell lines pathway feature vector by the trained model W yields a vector of cytokine mapping scores. Large positive mapping scores identify cytokines predicted to mediate innate drug resistance when they interact with the cancer cell line. (b) Multiplying a stromal cell lines cytokine feature vectors by the trained model W yields a vector of pathway mapping scores. Large mapping scores identify cancer pathways whose dysregulation is predicted to mediate resistance or sensitivity in the presence of the stromal cell line.

We trained a multi-task affinity regression model on co-culture data for 45 cancer cell lines and 23 stromal cell lines for 35 anti-cancer drugs from the Straussman et al. study [41], where for each drug, the multiple tasks consisted of the different drug dosages. We used rescue scores a quantification of cancer cell proliferation advantage conferred by co-culturing with stromal cells in the presence of drug as outputs. Our goal was to learn a model for cytokine-topathway interactions that would generalize to held-out cancer cells, so that we could, for example, predict the extent of stromal cell-mediated drug resistance (rescue score) for a test cancer cell line from its pathway score profile.

3.2.2 Multi-task affinity regression outperforms nearest neighbor methods for predicting rescue in co-culture experiments

In 10-fold cross-validation on held-out co-culture experiments, multi-task affinity regression strongly outperformed prediction based on stromal and cancer nearest neighbor methods, where the training co-culture experiment that is most similar to each test example on the basis of Euclidean distance in the cytokine or pathway feature space is considered the nearest neighbor ($P < 2.69 \times 10\text{-}14$ and $P < 4.77 \times 10\text{-}13$, respectively, Wilcoxon signed rank test; Fig. 1e,f). We also found that multi-task affinity regression strongly outperformed independent linear regressions across individual tasks in 10-fold cross-validation ($P < 8.72 \times 108$, Wilcoxon signed rank test). These results demonstrate the strong statistical performance of the tumor-stromal co-culture model learned with multi-task affinity regression.

In 10-fold cross-validation on held-out co-culture experiments, multi-task affinity regression strongly outperformed prediction based on stromal and cancer nearest neighbor methods, where the training co-culture experiment that is most similar to each test example on the basis of Euclidean distance in the cytokine or pathway feature space is considered the nearest neighbor ($P < 2.69 \times 10\text{-}14$ and $P < 4.77 \times 10\text{-}13$, respectively, Wilcoxon signed rank test; Fig. 3.4). We also found that multi-task affinity regression strongly outperformed independent linear regressions across individual tasks in 10-fold cross-validation ($P < 8.72 \times 10\text{-}8$, Wilcoxon signed rank test). These results demonstrate the strong statistical performance of the tumor-stromal co-culture model learned

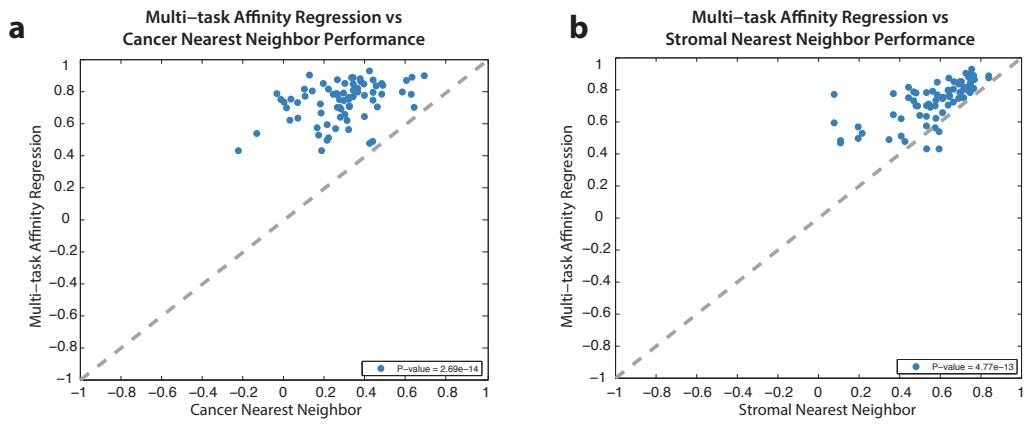

Figure 3.4: **Multi-task affinity regression outperforms nearest-neighbor methods in 10-fold cross-validation experiments.** Models were trained in multi-task fashion for 54 drug tasks corresponding to multiple dosage levels of 13 anti-cancer therapeutics; pan-EGFR inhibitor and pan-BRAF inhibitor models were also trained. Multi-task affinity regression strongly outperformed (a) cancer nearest neighbor ($P < 2.69 \times 10\text{-}14$, Wilcoxon signed rank test) and (b) stromal nearest neighbor ($P < 4.77 \times 10\text{-}13$, Wilcoxon signed rank test).

with multi-task affinity regression.

3.2.3 Multi-task affinity regression recovers HGF as a stromal factor that rescues melanoma cells treated with PLX4720

As the affinity regression model captures interaction information between cytokine features of the stromal cells and pathways of cancer cells, we next asked whether the trained model could identify which cytokines would elicit resistance in specific cancer cells. To achieve this, we trained a model W on all the

co-culture data for each drug and mapped each cancer cells pathway scores through the cytokine-pathway interaction matrix via CW to get a mapping score for each cytokine across cancer cell lines (Fig. 3.5). To assess the statistical significance of the mapping scores at each cytokine-cancer or pathway-stroma pairing, we trained 10,000 affinity regression models for different randomizations of the rescue scores, used the empirical null distribution of the mapping scores for each cytokine or pathway to define a nominal P value for the observed scores, and corrected for multiple tests for each cell line using the Benjamini-Hochberg procedure.

We then visualized these mapping scores for the PLX4720 model in a heatmap where significant cytokine-cancer cell interaction scores (FDR < 5%) are highlighted in orange (resistance) and blue (sensitivity) and (FDR <10%) are highlighted in dark orange (resistance) and dark blue (sensitivity) (Fig. 3.5). In the cytokine mapping, the cytokine-cancer interactions that received the highest scores in this heatmap consisted of a dark red column for HGF in the melanoma cell lines (Fig. 3.5). TNF-β was a second cytokine that was predicted by the model to elicit resistance in melanoma cells treated with PLX4720 (Fig. 3.5). In addition, there were distinct patterns of cytokine-cancer interactions for colorectal and melanoma cell lines, suggesting that colorectal and melanoma cells respond to different stromally-secreted cytokines (Fig. 3.5).

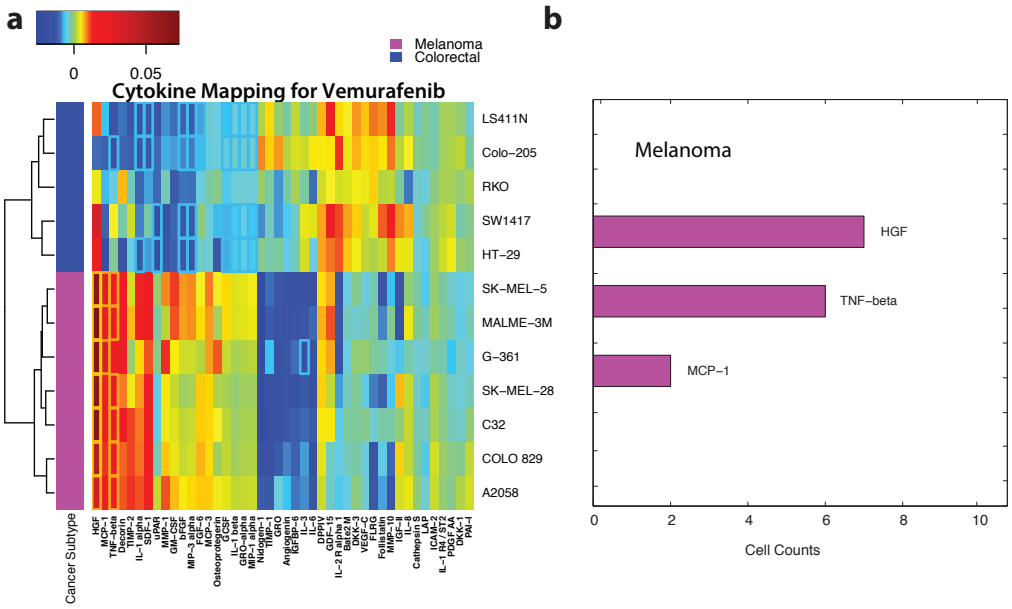

Figure 3.5: **Multi-task affinity regression identifies HGF as the main cytokine that elicits resistance in PLX4720 treated melanoma cell lines.** (a) The heatmap of the cytokine mapping scores for melanoma and colorectal cell lines using the vemurafenib model clusters cancer cell lines by subtype. Orange boxes indicate a cytokine-cancer cell line pair significant mapping scores relative to an empirical null model (FDR < 5%, see Methods) and dark orange boxes indicate a cytokine-cancer cell line pair significant mapping scores relative to an empirical null model (FDR < 10%, see Methods), identifying HGF as the main cytokine that elicits resistance in melanoma cell lines. (b) For each cytokine, the bar plots show the number of melanoma cell lines (magenta) for which the mapping score attained significance of (FDR < 10%). This analysis suggests that stromal-secreted HGF and TNF-beta frequently mediate resistance to vemurafenib in lung cancer cells.

3.2.4 Affinity regression implicates MET, MYC and PI3K pathways in stromal cell-mediated PLX4720 resistance

We next asked whether the trained model could infer which cancer cell pathways might receive signals from specific stromal cells to mediate drug resistance. We mapped each stromal cells cytokine expression levels through the cytokine-pathway interaction matrix, via WS to get a mapping score for each pathway across stromal cell lines (Fig. 3.6). We constructed a heatmap of these pathway-stromal mapping scores for PLX4720 treated co-cultures (Fig. 3.6), computed one-sided P values according to a null model similar to before, and highlighted scores associated with either drug resistance or sensitivity that passed an FDR threshold of 5% in orange (resistance) and blue (sensitivity) and FDR threshold of 10% in dark orange (resistance) and dark blue (sensitivity). Among significant stromal cell-pathway interactions (FDR < 10%) we found that the MYC pathway was associated with drug resistance in co-cultures with lung cancer cell lines. We found that the P53 regulation pathway was associated with drug sensitivity in co-cultures with lung and breast fibroblast cell lines. Notably, breast, lung, and skin stromal cells have different pathway interaction profiles based on the mapping score analysis (Fig. 3.6).

Finally, we interrogated the cytokine-pathway interaction matrix W for the PLX4720 affinity regression model. To assess the statistical significance of the interaction scores at each cytokine-pathway pairing, we again computed an empirical null distribution of scores for each cytokine-pathway pair, defined nominal P values for observed scores, and corrected for multiple tests using the Benjamini-Hochberg procedure for each cell line. We visualized these interaction scores in a heatmap of PLX4720 cytokine-pathway interactions, where sig-

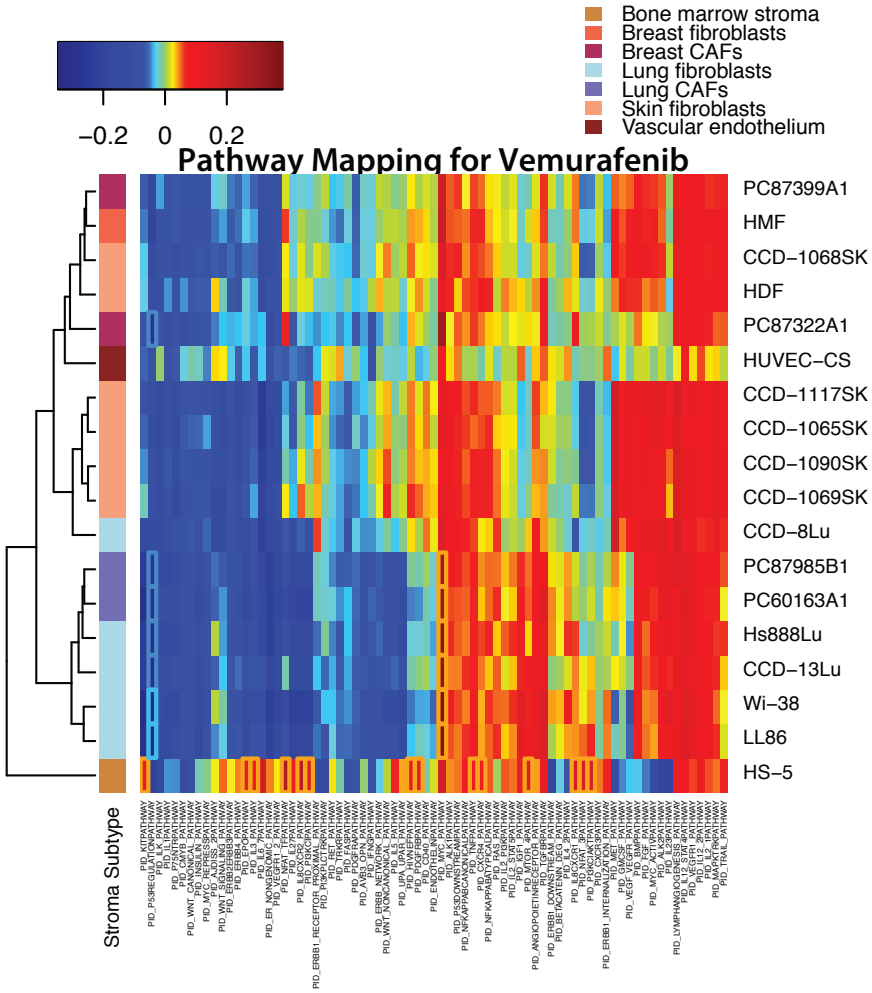

Figure 3.6: **Multi-task affinity regression implicates the MYC pathway in drug resistance and P53 regulation pathway in drug sensitivity in PLX4720 treated cell lines.** The heatmap of the pathway mapping scores for breast, skin, and lung fibroblasts using the vemurafenib model clusters stromal cell lines by their tissue of origin. Dark orange boxes indicate pairs whose mapping scores are significantly associated with drug resistance and dark blue boxes indicate pairs whose mapping scores are significantly associated with drug sensitivity (FDR <10%). This analysis implicates the MYC pathway in drug resistance and the P53 regulation pathway in drug sensitivity mediated by lung and breast fibroblasts.

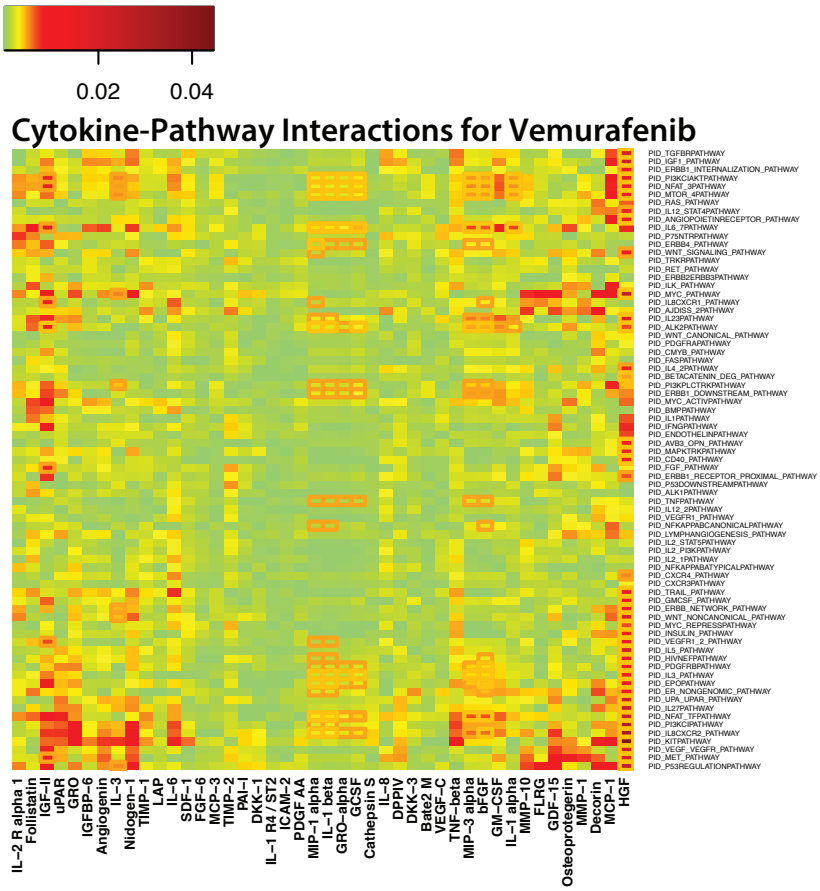

Figure 3.7: **Multi-task affinity regression identifies the HGF and MET pathway and HGF and PI3K and PI3K/AKT pathways in drug resistance against vemurafenib** The heatmap of absolute values of cytokine-pathway interactions in the model matrix W. Orange boxes indicate cytokine-pathway interactions that are significant relative to an empirical null model (FDR < 5%, see Methods) and dark orange boxes indicate (FDR < 10%, see Methods) and identify (HGF, MET pathway) and (HGF, PI3K pathway) and (HGF, PI3K/AKT pathway) as significant interactions.

nificant cytokine-pathway interactions (FDR < 5%) are highlighted in orange and (FDR < 10%) are highlighted in dark orange (Fig. 3.7). The cytokine-pathway interactions that received the highest scores were those that interacted with HGF, including the MET pathway, which is the receptor for HGF [10, 43], and the PI3K and PI3K/AKT pathway, which has been shown to be involved in HGF reactivation of oncogenic signaling in breast cancer cell lines treated with lapatinib and lung cancer cells treated with erlotinib [48].

3.2.5 Multi-task training learns a pan-EGFR inhibitor model of stromal-mediated resistance

After confirming that our model recovered the known biology of stromal-mediated resistance to PLX4720, we next turned to co-cultures treated with EGFR inhibitors. We trained separate multi-task affinity regression models on co-culture experiments with afatinib, canertinib, erlotinib, gefitinib, CL-387785, and lapatinib, where each model included data for multiple dosage levels of each drug. In addition, we trained a pan-EGFR model involving a subset of EGFR inhibitors with enriched activity against EGFR over other tyrosine kinases, including: afatinib, erlotinib, gefitinib, and CL-387785. Meanwhile canertinib, targeting EGFR, HER-2, and ErbB-4, and lapatinib, targeting EGFR and HER-2, were excluded as they inhibit multiple tyrosine kinases [5].

We then asked whether we could identify stromal-secreted cytokines contributing to drug resistance, as we did for the BRAF inhibitor PLX4720. To do this, we mapped the cancer pathways profile through the interaction matrix of the trained model (CW) to obtain a cytokine-cancer cell line mapping for indi-

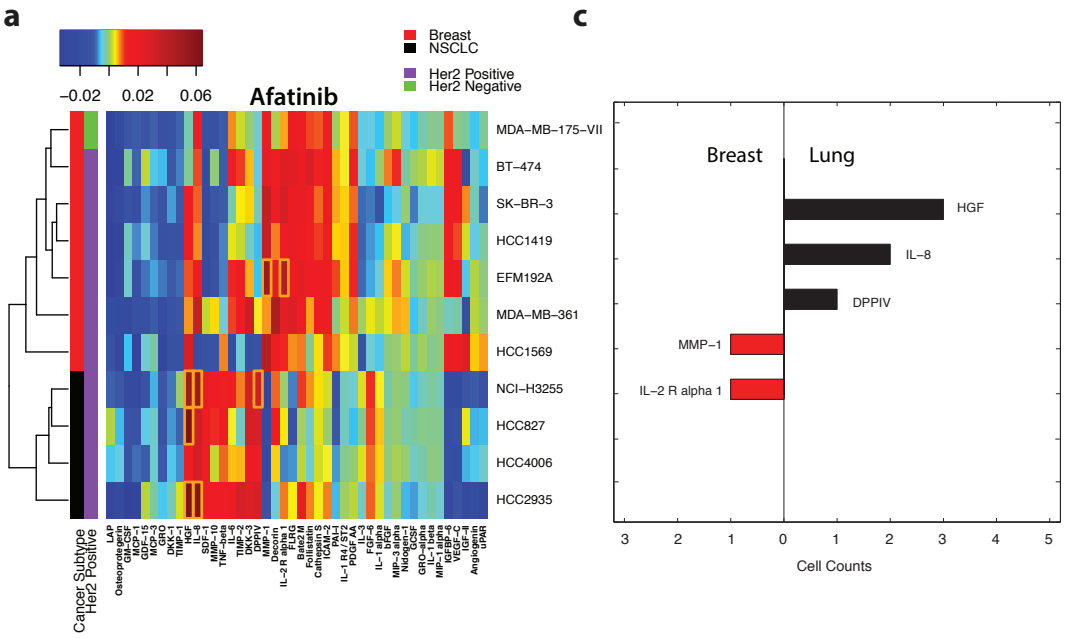

Figure 3.8: **Multi-task affinity regression identifies the HGF in drug resistance in afatnib treated co-cultures** (a) The heatmap shows cytokine mapping scores for breast cancer and non-small cell lung cancer (NSCLC) cells lines for the afatinib affinity regression model. Orange boxes indicate cytokine-cancer cell line pairs that are significantly associated with drug resistance relative to an empirical null model (FDR < 5%), while dark orange boxes indicate (FDR < 10%). (b) For each cytokine, the bar plots show the number of breast cancer cell lines (red) and the number of lung cancer cell lines (black) for which the mapping score attained significance. This analysis suggests that stromal-secreted HGF and IL-8 frequently mediate resistance to afatinib in lung cancer cells, but seldom in breast cancer cells.

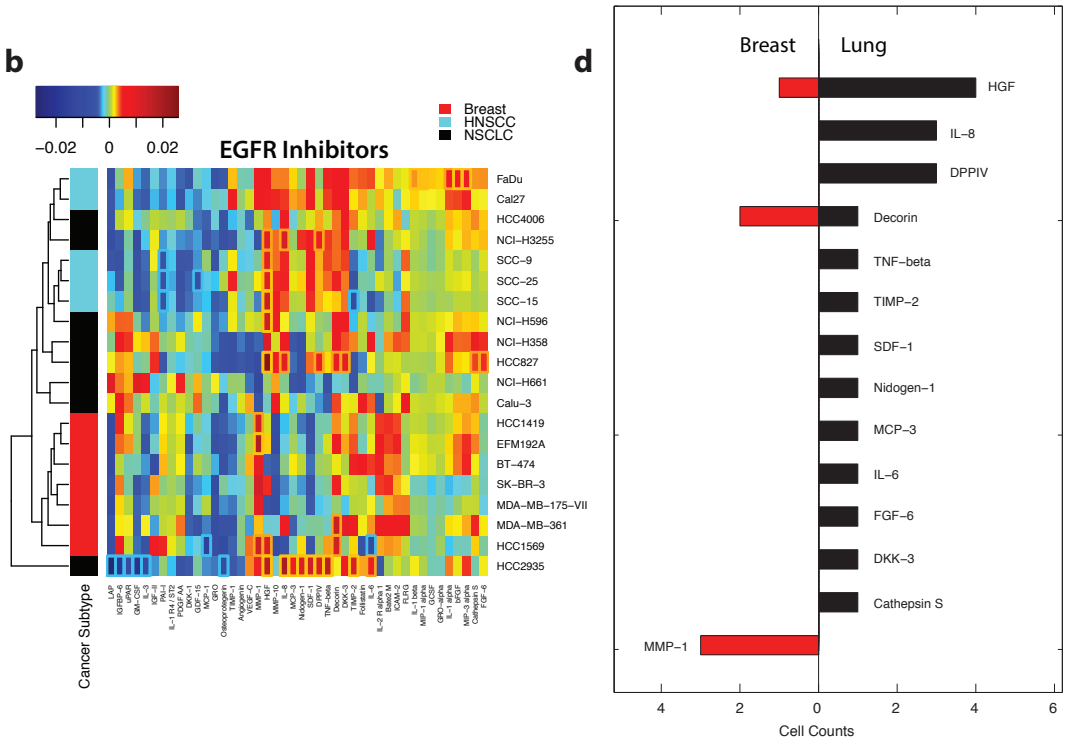

Figure 3.9: **Multi-task affinity regression identifies the HGF in drug resistance in pan-EGFR inhibitor model** (a) The heatmap shows cytokine mapping scores for breast cancer, NSCLC, and head and neck squamous cell carcinoma (HNSCC) cell lines for the pan-EGFR affinity regression model with multi-task training across gefitinib, afatinib, erlotinib, and CL-387785 (b) For each cytokine, the bar plots show the number of breast cancer cell lines (red) and the number of lung cancer cell lines (black) for which the mapping score attained significance for the pan-EGFR affinity regression model. The analysis suggests that HGF and IL-8 frequently mediate resistance to EGFR inhibitors in lung cancer cells, while MMP-1 mediates resistance in breast cancer cells.

vidual and the pan-EGFR inhibitor models. The cytokine-cancer cell mapping scores for the afatinib and pan-EGFR models, where we again used an empirical null model to assess the significance of high-scoring mapping scores and highlighted cytokine-cancer cell line interactions that satisfied an FDR < 5% threshold in orange and FDR < 10% in dark orange (Fig. 3.8, 3.9). Then we counted the number of breast cancer and lung cancer cell lines with a significant interaction for a cytokine (Fig 3.8, 3.9). Notably, this analysis suggested that HGF and IL-8 elicit resistance to afatinib and EGFR inhibitors in non-small cell lung cancer cells, but seldom in breast cancer cells (Fig. 3.8, 3.9).

3.2.6 Experimental validation confirms HGF as a novel stromal factor mediating resistance to afatinib and erlotinib in non-small cell lung cancer cells

Finally, to experimentally assess predictions from the affinity regression modeling of EGFR inhibitors, we tested the ability of HGF to rescue cancer cell proliferation in lung cancer cell lines treated with afatinib and erlotinib.

To do this, we cultured three NSCLC cell lines HCC4006, HCC2935, and HCC827 all of which were also included in the Straussman et al. co-culture data sets and performed dose response experiments to afatinib and erlotinib in the presence or absence of HGF (Fig. 3.10). In all cases, we confirmed that HGF conferred rescue to the afatinib and erlotinib treated NSCLC cell lines.

Dose response curves in the presence of drug only (blue) or drug with HGF (red) as well as the relative cancer proliferation at the dose with maximal rescue

Significance of HGF mediated resistance to EGFR inhibitors in Non-Small Cell Lung Cancer

	HCC4006	HCC2935	HCC827
Afatinib	8.3e-5	2.0e-4	8.7e-3
Erlotinib	7.1e-4	3.3e-5	1.0e-4

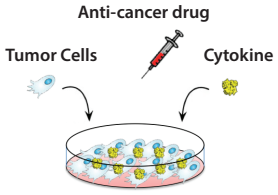

Figure 3.10: **Summary of HGF mediated drug resistance non-small cell lung cancer cell lines** Summary of drug resistance experiments comparing cancer cell line proliferation under treatment with two EGFR inhibitors, afatinib and erlotinib, across 3 NSCLC cell lines with and without HGF. The P value indicates the most significant increase in cancer cell counts over the drug dose response experiments in drug+HGF versus drug only conditions.

are shown from specific examples (dose response for HCC4006 with afatinib concentrations 0.49 nM 2000 nM and relative abundance at 31 nM; HCC2935 with afatinib concentrations 31 nM 2000 nM and at 32 nM; HCC4006 with erlotinib concentrations 6.4 nM 20000 nM and at 32 nM; HCC2935 with erlotinib concentrations 0.05 nM 20000 nM and at 32 nM), Fig. 3.11.

3.3 Methods

3.3.1 Rescue score calculation

We adapted the rescue score associated with drugs in tumor-stromal co-culture experiments as previously defined by Straussman et al. [41]. The original study provided data on cancer cell proliferation under drug treatment as the normal-

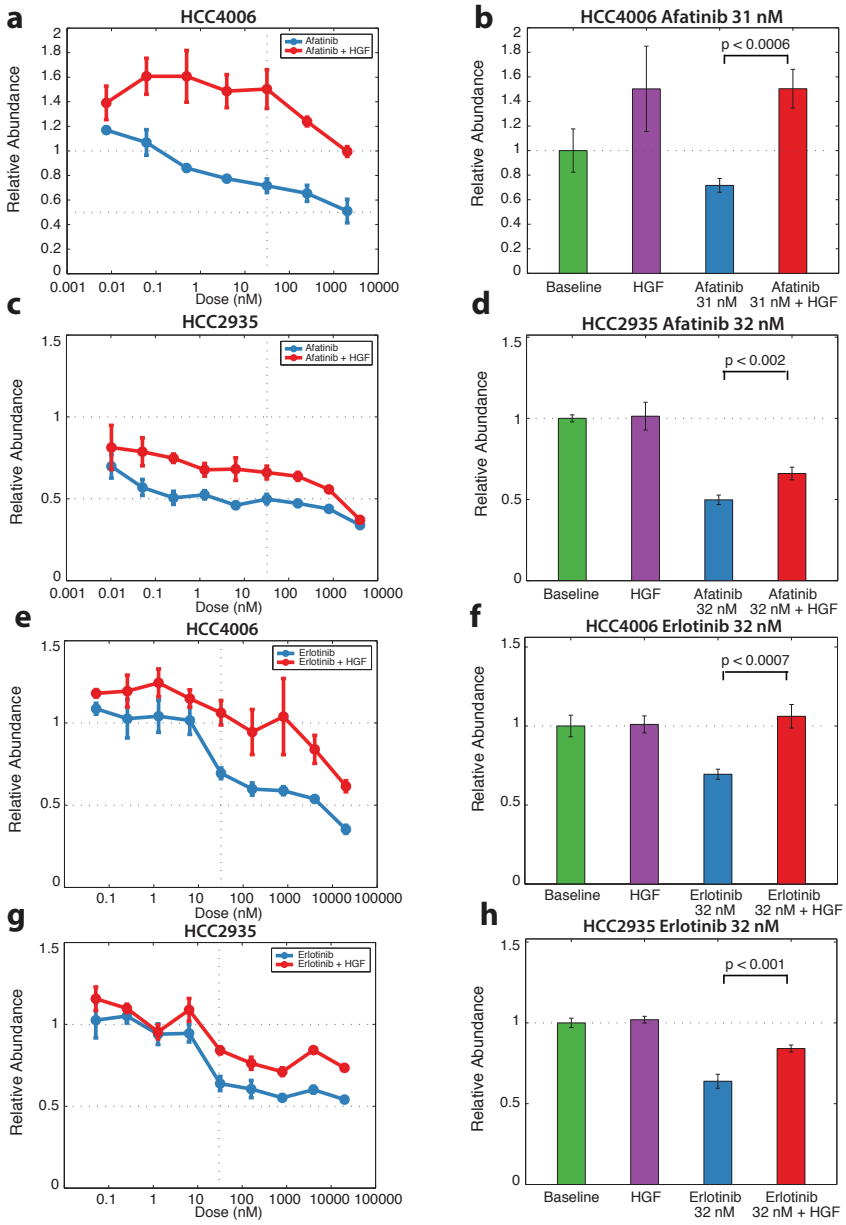

Figure 3.11: **Dose response curves and relative abundance of cancer proliferation at a specific drug dose for HGF mediated drug resistance in non-small cell lung cancer cell lines** (b,d,f,h) Dose response curves for drug+HGF (red) versus drug only (blue) for specific NSCLC cell lines, plotting relative abundance of cancer cells as a function of drug dose with HGF concentration fixed. Experiments are performed in triplicate. (c,e,g,i) Relative abundance of cancer cells at a specific drug dose (as shown), showing baseline, HGF only, drug only, and HGF values.

ized GFP cell count of cancer cells in monoculture (M) and co-culture with stromal cells (C). We calculated a proliferation pseudocount p that was approximately 0.65 and saw that the choice of pseudocounts did not largely affect the performance of the model. We defined the rescue score R as the difference of log transformed. We computed Z-scores from the resulting values to get the new rescue scores:

$$R = log2(M + p) - log2(C + p). \tag{3.2}$$

3.3.2 Cytokine array processing

We obtained cytokine expression levels from the Human Cytokine Array G4000 (RayBio, AAH-CYT-G4000-8) from the Straussman et al. study [41]. We shifted the entire distribution of cytokine expression levels by adding the minimum value plus 1, and log10-transformed the resulting values. Then we converted the cytokine expression values to Z-scores. Finally, we filtered out cytokines whose maximum Z-score was less than 0.75; the Z-scores for the remaining cytokines gave the stromal input matrix used in training the model.

3.3.3 Pathway scores

We performed standard RMA normalization on microarray gene expression data of 1036 cancer cell lines available through the Cancer Cell Line Encyclopedia (CCLE). We curated 75 relevant pathways from the Pathway Interaction Database available through the GSEA web resource, selecting signal transduction and oncogenic pathways. For the 39 cancer cell lines in the co-culture exper-

iments with expression data in CCLE, we calculated pathway scores as follows. For each pathway, we restricted the expression matrix of all 1036 CCLE cancer cell lines to the genes in the pathway, performed PCA on this reduced expression matrix, and took the coordinate of each cell line relative to the first principal component as its pathway score. Since the sign of the coordinate is arbitrary, we took the positive direction to be the one where the larger number of genes in the pathway had a positive sign. For the final pathway score input matrix, the vector of pathway scores for each cancer cell line was unit normalized.

3.3.4 Multi-task affinity regression

We modeled the interaction between cancer and stromal cells and their molecular components in co-culture under drug treatment using a bilinear regression method called affinity regression [40]. In this approach, we try to model how the stromal cells signal to the cancer cells through secreted cytokines and how cancer cells receive the signals through activated signaling pathways. Affinity regression predicts the rescue scores based on interactions between the cancer cell line pathway scores and stromal cell line cytokine expression levels. Since co-culture experiments treated with different drug dosages had been performed, we jointly learned from multiple dosage experiments using a multi-task version of affinity regression.

$$argmin_W \sum_{t=1}^{T} ||vec(Y_t) - (S_t \otimes C_t)vec(W_t)||_2^2 + \rho_1 \sum_{t=1}^{T} ||vec(W_t) - W_o||_2^2 + \rho_2 \sum_{t=1}^{T} ||vec(W_t)||_2^2 \tag{3.3}$$

$$W_o = \frac{1}{T} \sum_{s=1}^{T} W_s \tag{3.4}$$

A reconstruction of the optimal common task model is given by

$$W_o^* = \frac{\rho_1}{\rho_1 + \rho_2} \frac{1}{T} \sum_{t=1}^{T} W_t^* \tag{3.5}$$

3.3.5 Parameter optimization

We trained multi-task affinity regression for 13 targeted drug therapies and 2 larger drug groups consisting of EGFR and BRAF inhibitors, comprising a total of 54 different co-culture experiments across the different training sets. For each model, we performed a parameter grid search for the optimal parameters ρ_1 and ρ_2, varying ρ_1 from 0.00001 to 10000 and ρ_2 from 0.00001 to 10000. We narrowed down the parameter space for both ρ_1 and ρ_2 to a range of 9 parameter values and performed nested 10-fold cross-validation.

We performed a nested 10-fold cross-validation of the multi-task affinity regression. For each fold of the outer 10-fold cross-validation, we performed a grid search across 9x9 parameter pairs (ρ_1 and ρ_2). We took the median of median parameter values to obtain our optimal parameter pair from nested 10-fold cross-validation. Using this parameter choice we calculated the mean task-specific Spearman correlation performance for all 54 drug dosages.

3.3.6 Comparison of multi-task with single-task affinity regression

We compared the performance of our task specific predictions of our multi-task affinity regression with the performance of least squares affinity regression on the tasks or 54 drug dosages as separate regressions. We constrained the least squares optimization with a L_2 regularizer where we varied the regression parameter ρ from 0.0001 to 100 incrementally by orders of 10 magnitudes. We performed a nested 10-fold cross-validation of the least squares affinity regression and obtain a mean Spearman correlation as a measure of performance in the outer 10-fold cross-validation for each of 54 drug dosages across 13 drugs and 2 larger drug groups.

3.3.7 Comparison of multi-task affinity regression with nearest neighbor methods

The stromal nearest neighbor algorithm finds the predicted rescue score for a held-out drug co-culture experiment by reporting the rescue score of the training co-culture experiment with the stromal cell line that is the closest in Euclidean distance in the stromal feature space and with the same cancer cell line. Similarly, the cancer nearest neighbor algorithm finds the predicted rescue score by reporting the rescue score of the training co-culture experiment with the cancer cell line that is the closest cancer cell line in cancer feature space and with the same stromal cell line. We compared the Spearman correlation of the predictions in the 54 drug/dosage data sets to the experimental rescue values in multi-task

affinity regression to cancer and stromal nearest neighbor algorithms.

For each comparison of the Spearman correlations between the multi-task affinity regression and nearest neighbor from 54 drug/dosage experiments, we performed a Wilcoxon signed rank test to see if multi-task affinity regression outperforms nearest neighbor in Spearman correlation.

3.3.8 Empirical null models

For assessing the significance of cytokine-cancer cell line mapping scores, we generated 10,000 randomized data sets where we permuted the rescue scores of the response variable Y. We permuted the rescue scores for each stromal cell line and then we permuted the stromal cell lines. Then we trained 10,000 multi-task affinity regression models and mapped the W interaction matrices onto the cancer matrix C. For each cytokine-cancer mapped value, its empirical p-value is the fraction of randomly generated mapping values more extreme than that value out of the distribution of randomly generated mapping values for that cytokine-cancer pair. We obtained p-values for both the left-hand sides and right hand-sides of the distributions. The left hand side tested for a cytokine-cancer or pathway-stromal interactions sensitivity to the drug and the right hand side tested for resistance to the drug. We corrected for FDR in multiple hypothesis testing of cytokine-cancer mapping values for each cell line separately using the Benjamini-Hochberg procedure. Pathway-stromal mapping P values were computed in an analogous fashion, again with Benjamini-Hochberg FDR correction applied for each stromal cell line separately.

For assessing the significance of cytokine-pathway interaction weights in the

W matrix, we generated random 10,000 data sets where we permuted the rescue scores in the response variable Y. We permuted the rescue scores for each stromal cell line and then we permuted the stromal cell lines. Then we ran 10,000 multi-task affinity regression models and obtained the W interaction matrices. For each cytokine-pathway interaction value, its empirical p-value is the number of randomly generated interaction values found more extreme than that value over the total number of randomly generated interaction values for that cytokine-pathway interaction.

We performed a Benjamini-Hochberg FDR correction for the set of P values associated with each cytokine separately.

3.3.9 Cells, inhibitors, and cytokines

HCC827, HCC4006 and HCC2935 cells were a kind gift from the lab of Marc Ladanyi (MSKCC). Cells were maintained in DMEM with 10% fetal bovine serum, penicillin and streptomycin. HGF-1 (peprotech) was dissolved in PBS+0.5% BSA was used at 50 ng/ml. Erlotinib and afatinib were purchased from Sellechem and used at an assay dependent concentration as indicated. For drug and cytokine treatment, 5,000 cells were plated per well in a 96 well plate with DMEM+10%FBS. Growth was measured in triplicate following 72 hours of drug/cytokine treatment using an MTT assay per the manufacturers instructions (Roche). The experimental validation was performed by Dr. Robert L. Bowman in Dr. Johanna Joyce's lab at Memorial Sloan Kettering cancer center.

3.4 Conclusion

We have presented a new supervised learning strategy for modeling cancer-stromal cell paracrine signaling from co-culture anti-cancer drug sensitivity experiments. Using an expression-based pathway feature representation for cancer cells and a cytokine expression level representation for stromal cells, we trained affinity regression models to predict stromal-mediated rescue scores of cancer cells for each drug; we employed a multi-task strategy to share information across models for different dosages of the same drug or different drugs of the same class. As a bilinear regression model, affinity regression allows us to define feature space mappings using the trained model. The mappings identified the cytokines that are most strongly associated with resistance/sensitivity for each cancer cell line, as well as the cancer cell pathways that appear to mediate resistance/sensitivity for each stromal cell line. Through an empirical null model, we assigned statistical significance to these key predicted features. This analysis recovered the published finding that stromal-derived HGF mediates resistance to BRAF inhibitors in melanoma cell lines. Moreover, our affinity regression analysis predicted that HGF would mediate resistance to EGFR inhibitors in lung cancer cell lines. Follow-up cell culture experiments with afatinib and erlonitib confirmed this clinically relevant prediction.

The largest limitation of our study is the small training data set size. Ideally, we would use on the order of 100 stromal and cancer cell lines with a full matrix of co-culture experiments as training data for each drug/dosage; in practice, the co-culture matrix used consisted of 10-20 cancer cell lines by a similar number of stromal cell lines, and a different matrix of co-cultures was assayed for each drug [41]. While our multi-task strategy helps improve model

accuracy in this low training data setting, it cannot fully address the fact that we are not adequately sampling the space of cancer-stromal interactions. As a secondary issue, the mRNA expression signature representation of cancer was also predicated on data availability; phosphoproteomic data might be more directly relevant as a representation of active signaling pathways in cancer cells, but while data resources based on technologies like reverse-phase protein array [25] and mass spectrometry [31] are increasing, the overlap with the cancer cell lines in our study remains limited. Nevertheless, the current work provides an important proof-of-principle that supervised learning can indeed derive novel findings from co-culture drug sensitivity data sets, providing a path forward for future larger-scale studies.

3.5 Comparison of multi-task affinity regression to multi-task pairwise SVM

3.5.1 Optimization Problem for Pairwise SVM

Support Vector Machines are classification algorithms that learn a linear decision rule based on maximizing a margin [3]. Different mappings $x \mapsto \Phi(x) \in H$ construct different SVMs. The mapping $\Phi(\cdot)$ is implicitly performed by a kernel function $K(\cdot, \cdot)$ which defines an inner product in H. The decision function given by an SVM can be described by:

$$f(x) = w \cdot \Phi(x) + b = \sum_i \alpha_i y_i K(x_i, x) + b$$

(3.6)

The optimal hyperplane is the one with maximal distance (in H space) to the closest image $\Phi(x_i)$ from the training data (called the maximal margin). This reduces to the following dual optimization problem:

$$\arg \alpha \max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

(3.7)

subject to:

$$0 \leq \alpha_i \leq C, \sum_{i=1}^{m} \alpha_i y_i = 0, i = \{1, ..., m\}$$

In affinity regression, our examples consist of each (i, j) tumor-stromal pair and our regression features consist of the pairwise products of the features of each input matrix given by $v_j \otimes u_i$. We set up an equivalent SVM formulation in this feature space. For this purpose, we discretize our response variable Y such that $y_{ij} \in \{+1, -1\}$. Our w is now the normal to the hyperplane that divides the training examples with the maximum margin.

$$Y = w \cdot X \quad \Longrightarrow \quad y_{i,j} = w \cdot (v_j \otimes u_i)$$

(3.8)

To obtain the w with the maximum margin we solve this optimization problem:

$$\arg w, \xi \min \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{m} \sum_{j=1}^{n} \xi_{i,j} \tag{3.9}$$

subject to:

$$y_{i,j}(w \cdot (v_j \otimes u_i)) \geq 1 - \xi_{i,j}, \xi_{i,j} \geq 0, i = \{1, ..., m\}, j = \{1, ..., n\}$$

It can be shown that the w is a linear combination of the feature vectors of the examples.

$$w = \sum_{s,t} \alpha_{s,t} y_{s,t} (v_s \otimes u_t) \tag{3.10}$$

substituting the for w in the expression for $y_{i,j}$ (Equation 4.3) we get:

$$y_{i,j} = \sum_{s,t} \alpha_{s,t} y_{s,t} (v_s \otimes u_t)(v_j \otimes u_i)$$

(3.11)

We use the tensor kernel:

$$K((v_i, u_j), (u_t, v_s)) = (v_j \otimes u_i)(v_s \otimes u_t)$$

(3.12)

Furthermore if we distribute the terms v_i and u_s through the kronecker product, we get:

$$y_{i,j} = \sum_{s,t} \alpha_{s,t} y_{s,t} (v_j \otimes u_i)(v_s \otimes u_t) = \sum_{s,t} \alpha_{s,t} y_{s,t} (v_j \cdot u_s) \cdot (v_i \cdot u_t)$$

(3.13)

The affinity regression model can be formulated as a pairwise SVM where the tasks are symmetric and can either be on the cancer side or on the stromal side.

$$K((u_i, v_j), (u_t, v_s)) = K_v(v_j, v_s) \cdot K_u(u_i, u_t)$$

(3.14)

where:

$$K_v(v_j, v_s) = v_j \cdot v_s$$
$$K_u(u_i, u_t) = u_i \cdot u_t$$

(3.15)

K_v is the Kernel for similarities between stromal cell lines.

K_u is the Kernel for the similarities between cancer cell lines.

3.5.2 Primal Optimization Problem for Multi-task Pairwise SVM

To multi-task the pairwise SVM, we turn to the multi-task formulation in the Evgeniou and Pontil paper [46]. We solve this optimization problem:

$$\arg \mathrm{w}_t, \xi_{i,j,t} \min \rho_1 \sum_{t=1}^{T} \parallel w_t \parallel^2 + \rho_2 \sum_{t=1}^{T} \parallel w_t - \frac{1}{T} \sum_{s=1}^{T} w_s \parallel^2 + \sum_{t=1}^{T} \sum_{i=1}^{m} \sum_{j=1}^{n} \xi_{i,j,t}$$

(3.16)

subject to:

$$y_{i,j,t}(w_t \cdot (v_{j,t} \otimes u_{i,t})) \geq 1 - \xi_{i,j,t}, \xi_{i,j,t} \geq 0, i = \{1, ..., m\}, j = \{1, ..., n\}, t = \{1, ..., T\}$$

3.5.3 Dual Optimization Problem for Multi-task Pairwise SVM

Let

$$\text{C:}=\frac{1}{2\cdot\rho_1+\rho_2},\mu:=\frac{T\rho_1}{\rho_2}, \quad (3.17)$$

and define the kernel

$$\mathrm{K}_{s,t}((v_j\otimes u_i),(v_l\otimes u_k)):=(\frac{1}{\mu}+\delta_{s,t})(v_j\otimes u_i)\cdot(v_l\otimes u_k) \quad (3.18)$$

where $i,k\in\{1,2,...,m\}$, $j,l\in\{1,2,...,n\}$, and $s,t\in\{1,2,...,T\}$

The dual problem of (Equation 4. 11) is given by

$$\arg\alpha_{i,j,t}\max\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{t=1}^{T}\alpha_{i,j,t}-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{s=1}^{T}\sum_{k=1}^{m}\sum_{l=1}^{n}\sum_{t=1}^{T}\alpha_{i,j,s}y_{i,j,s}\alpha_{k,l,t}y_{k,l,t}K_{s,t}((v_{j,s}\otimes u_{i,s}),(v_{l,t}\otimes u_{k,t})) \quad (3.19)$$

subject, for all $i,k\in\{1,2,...,m\}$, $j,l\in\{1,2,...,n\}$, and $s,t\in\{1,2,...,T\}$, to the constraints that the constraint that $0\le\alpha_{i,j,t}\le C$.

3.6 Performance Comparison of multi-task affinity regression to multi-task pairwise SVM

I compared the performance of multi-task affinity regression with the multi-task pairwise SVM using 10-fold cross-validation on the co-culture drug screen data set. I created labels from the rescue scores by taking rescue scores > 0.6 as my positive labels and rescue scores < 0.2 as my negative labels. Here a positive label means there was drug resistance conferred to the cancer cells by the stromal cells in co-culture and a negative label means no drug resistance was conferred. I ran the multi-task pairwise SVM models across co-cultures treated with different drug dosages for 13 anti-cancer drugs and across different drugs for two drug groups: the BRAF/MEK and EGFR inhibitors using 10-fold cross-validation. For comparison with affinity regression, I calculated an AUC for the corresponding multi-task affinity regression models by converting the continuous rescue scores to positive and negative labels as described above and then calculated the AUC for the predicted rescue scores from affinity regression. Here I show the comparison in performance for the two methods for some example drugs: PLX4720, afatinib, erlotinib, and gefinitib. Finally, I show the overall comparison of AUCs for all 13 drugs and the two drug groups. Multi-task affinity regression outperforms multi-task pairwise SVM ($P < 3.05 \times 10\text{-}4$, Wilcoxon signed rank test). The AUC for both methods were high. If we adopt a nonlinear Gaussian or polynomial kernel the pairwise SVM could possibly outperform affinity regression.

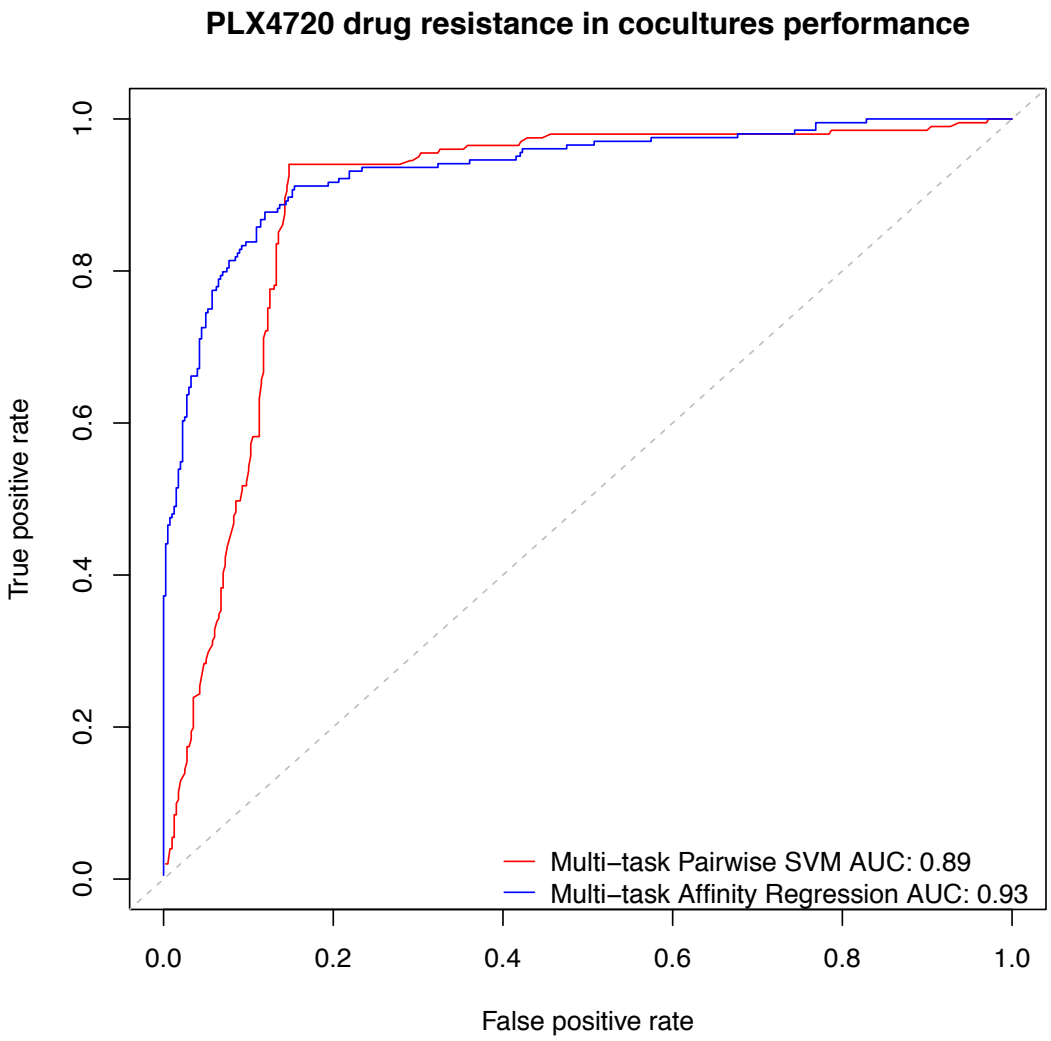

Figure 3.12: **Comparison of multi-task affinity regression with multi-task pairwise SVM in PLX4720 drug resistance** Multi-task affinity regression (AUC = 0.93) outperforms multi-task pairwise SVM (AUC = 0.89). Models were assessed using 10-fold cross-validation.

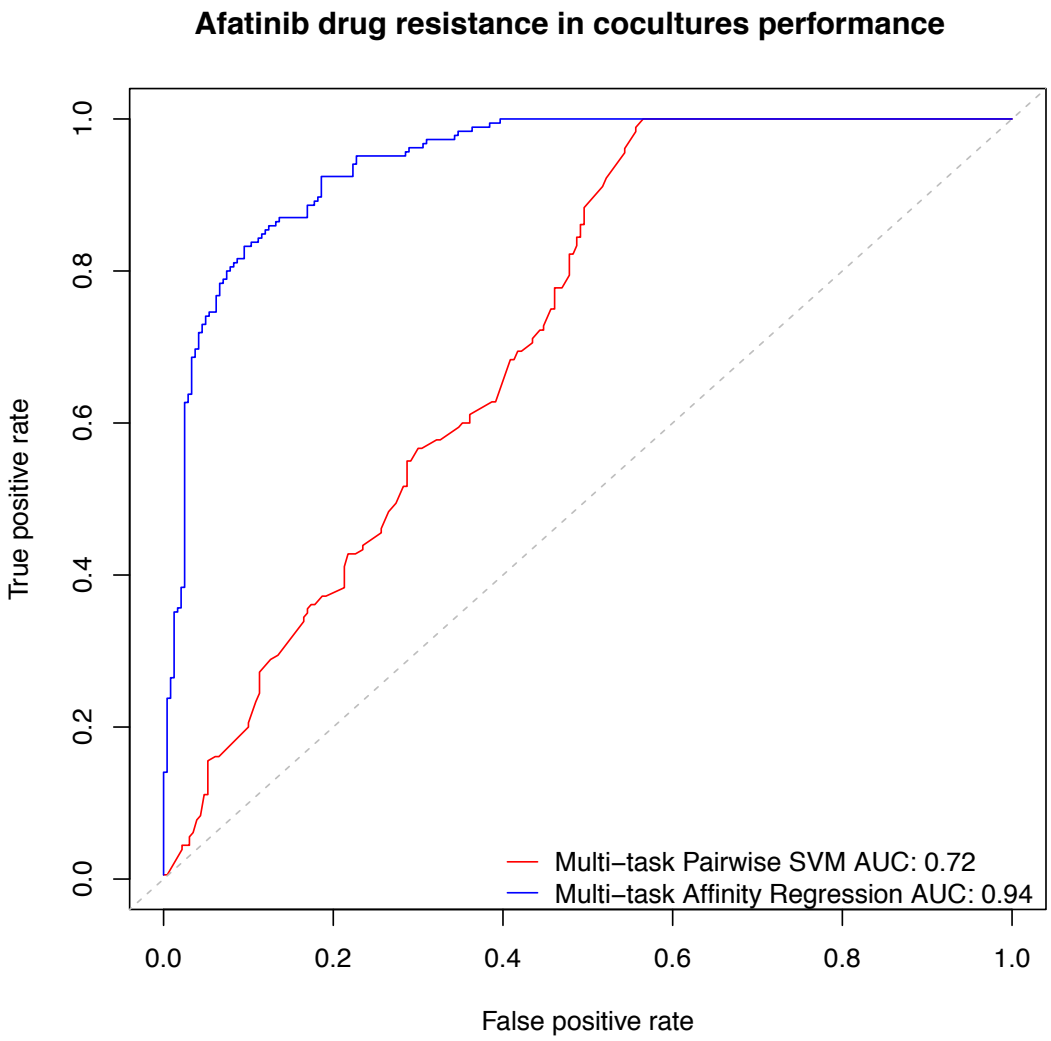

Figure 3.13: **Comparison of multi-task affinity regression with multi-task pairwise SVM in afatinib drug resistance** Multi-task affinity regression (AUC = 0.94) outperforms multi-task pairwise SVM (AUC = 0.72). Models were assessed using 10-fold cross-validation.

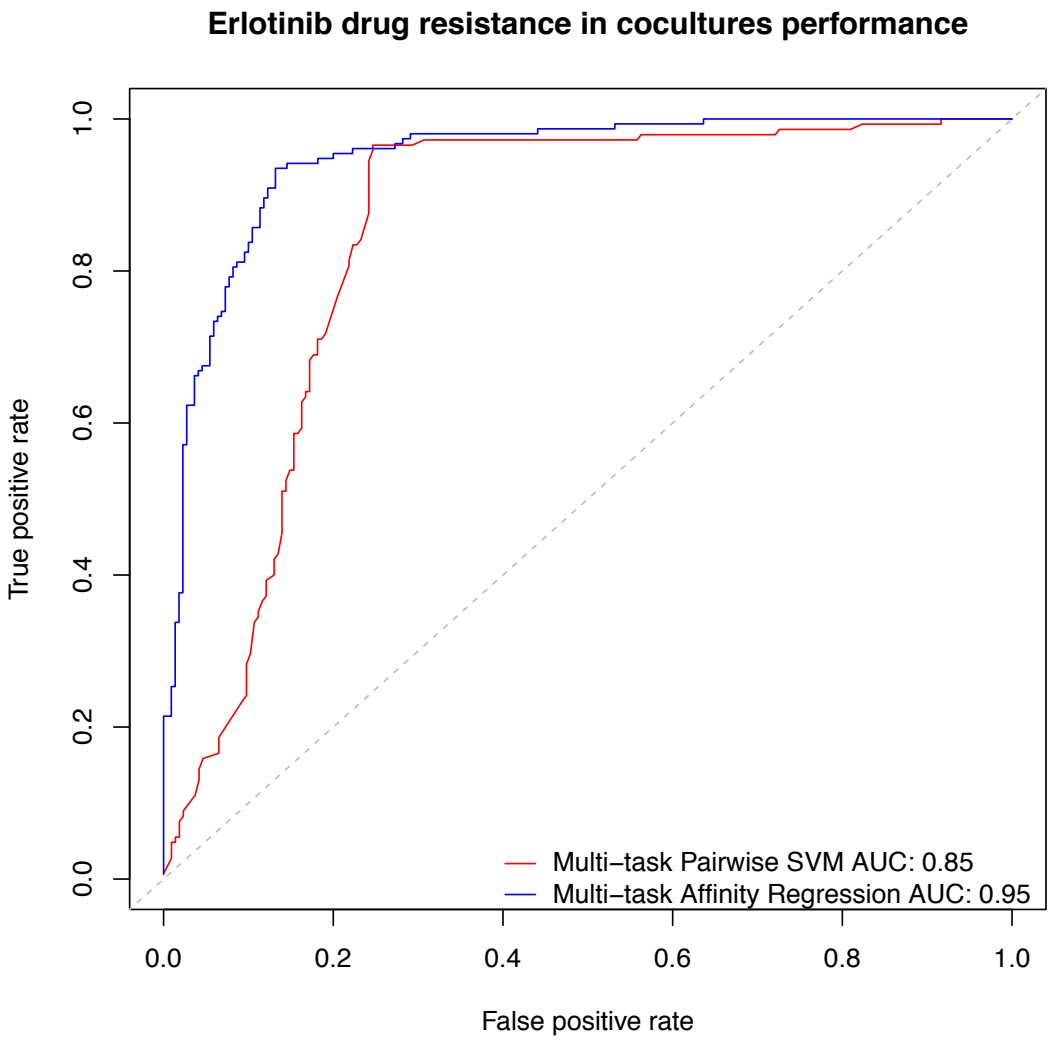

Figure 3.14: **Comparison of multi-task affinity regression with multi-task pairwise SVM in erlotinib drug resistance** Multi-task affinity regression (AUC = 0.95) outperforms multi-task pairwise SVM (AUC = 0.85). Models were assessed using 10-fold cross-validation.

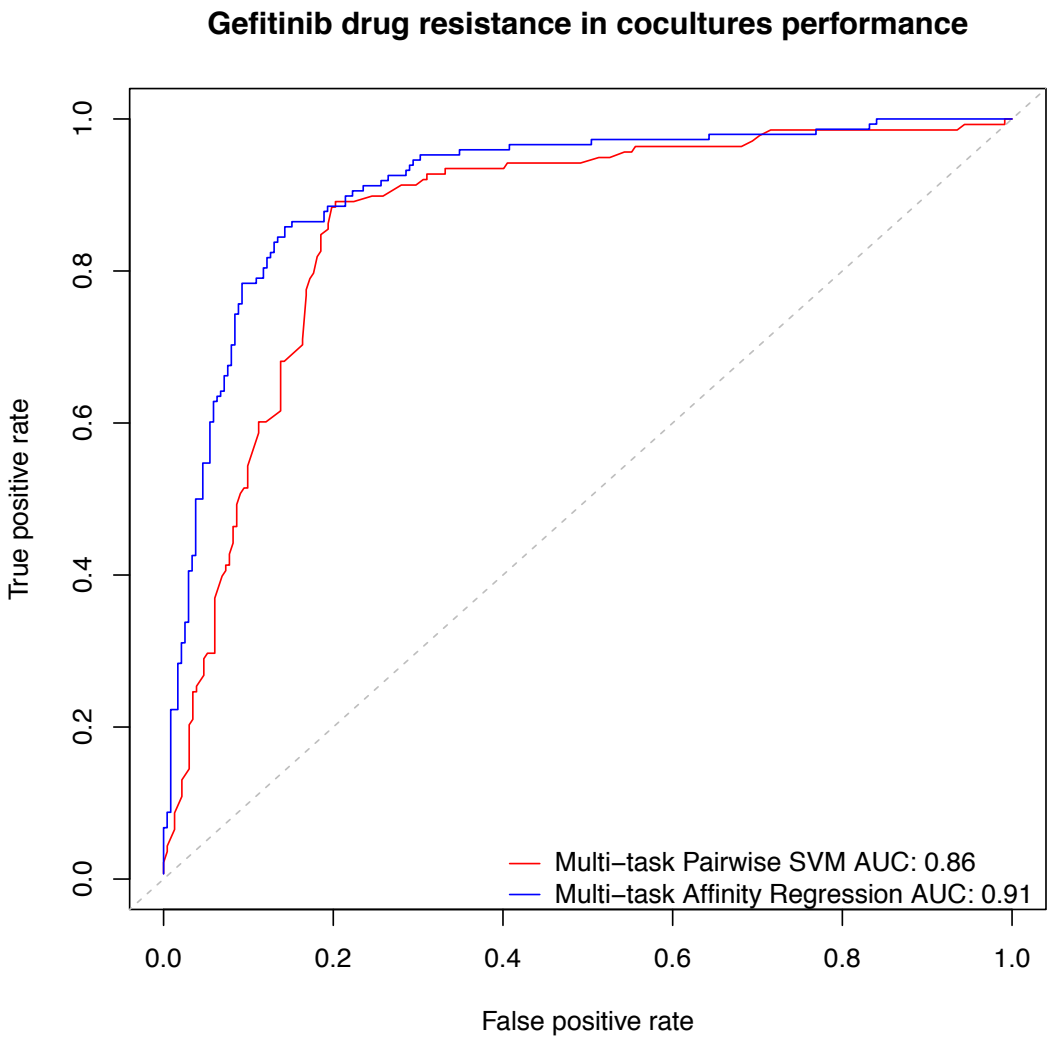

Figure 3.15: **Comparison of multi-task affinity regression with multi-task pairwise SVM in gefitinib drug resistance** Multi-task affinity regression (AUC = 0.91) outperforms multi-task pairwise SVM (AUC = 0.86). Models were assessed using 10-fold cross-validation.

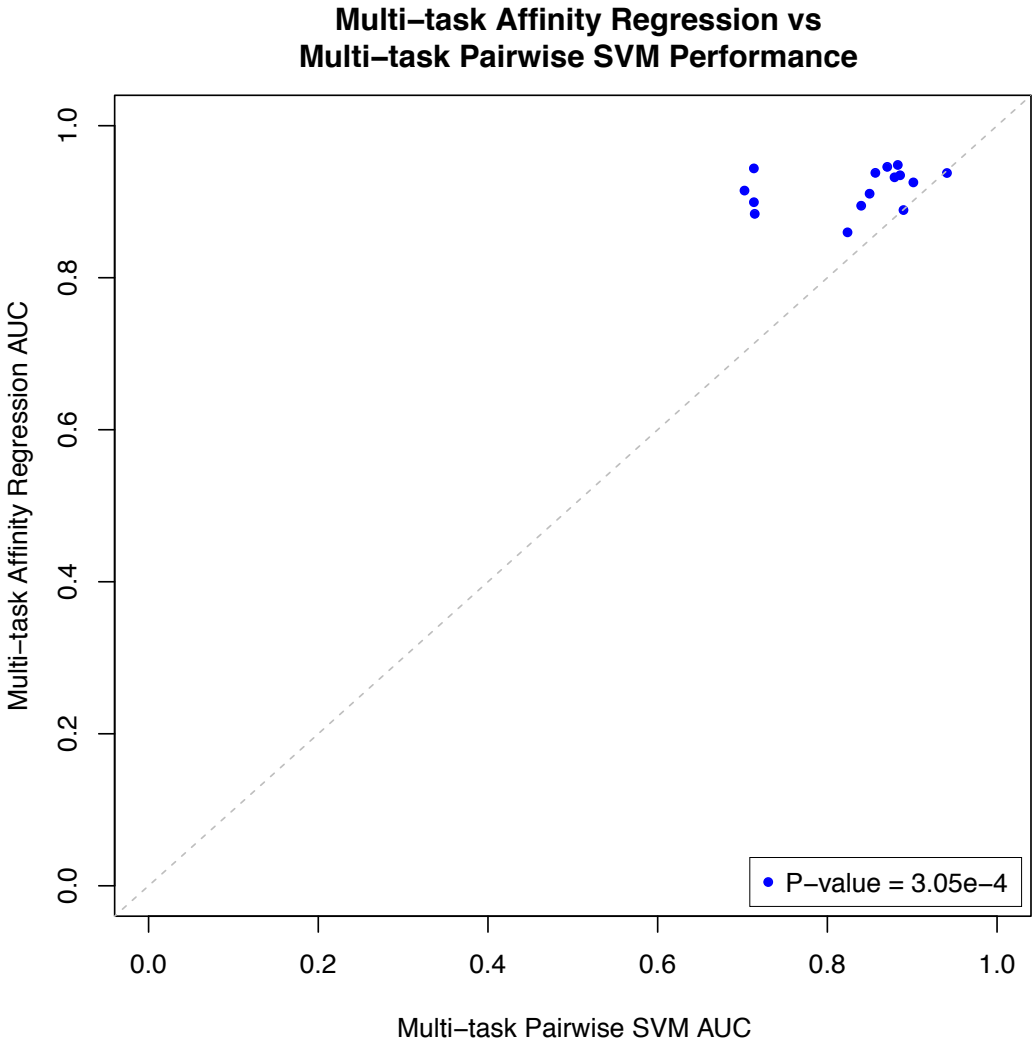

Figure 3.16: **Comparison of multi-task affinity regression with multi-task pairwise SVM AUC performance in 13 drugs and 2 drug groups.** Multi-task affinity regression outperforms multi-task pairwise SVM in AUC ($P < 3.05 \times 10\text{-}4$, Wilcoxon signed rank test). Models were assessed using 10-fold cross-validation.

CHAPTER 4

APPLICATION OF AFFINITY REGRESSION TO PBM AND RNACOMPETE DATA

4.1 Application of affinity regression to PBM and RNAcompete data

4.1.1 Affinity regression modeling of TF/RBP and DNA/RNA interactions in PBM and RNAcompete data

We used affinity regression in another context to analyze PBM and RNAcompete data sets. In PBM experiments, the DNA-binding preferences of an individual fluorescently tagged TF are measured using a universal array of >40K double-stranded DNA probes [12]. In the RNAcompete assay, the binding affinity of an RBP is measured against >200K single-stranded RNA probes [13, 14]. In our 2015 affinity regression paper [40], we used affinity regression to learn the DNA or RNA recognition code for families of TFs or RBPs directly from the protein sequence and probe-level binding data from PBM or RNAcompete experiments. In this context, affinity regression trains on PBM or RNAcompete experiments to learn an interaction model between proteins and nucleic acids. Learning from PBM or RNAcompete data for diverse TFs and RBPs, the affinity regression model can predict the binding affinities of held-out proteins and identify key DNA/RNA-binding residues.

In this project, I prepared the input data and generated and analyzed the

motifs for affinity regression models trained on RNAcompete data and generated motifs for affinity regression models for PBM data. I obtained the amino acid sequences of protein binding domains of 207 RBPs. I computed the amino acid 4-mer count matrix for the protein binding domain input sequences of the RBPs. I also created the nucleotide 5-mer count matrix for the nucleic acid input sequences of the RNA probes. I then extracted and normalized the matrix of binding affinities of the RBPs to the RNA probes. These data were then input into the affinity regression algorithm to learn the binding affinities of the RBPs to the RNA probes as a function of the interactions between the K-mer features of the protein domain sequences and nucleotide k-mer features of the RNA probes.

Binding models can be trained either from probe level measurements or 8-mer/7-mer summarizations of the PBM/RNAcompete probe-level data. We also trained two affinity regression models from the 7-mer Z-score summarizations of the probe-level data using Z-scores as reported on the cisBP-RNA website and the from the PBM data set. After the modeling was done for the Z-score models, I plotted an example of the experimental and affinity regression predicted Z-scores for SNAPOd2T00005194001, a mouse homeodomain from the PBM data set, showing the top 100 affinity regression predicted 8-mers versus the top 100 experimental 8-mers and the enrichment of top experimental 8-mers in the predicted set. I generated motifs from the DNA/RNA-binding residues of probes that were associated with high Z-scores from the Z-score models. For the RBP Z-score model, I evaluated how well affinity regression was able to recover these motifs by comparing the motifs obtained from affinity regression against the motifs obtained from a nearest neighbor competitor algorithm and motifs obtained directly from the RNAcompete assay, which we considered as

the ground truth. I compared the Kullback-Leibler divergence (D_{KL}) of the affinity regression motifs with the D_{KL} of nearest neighbor motifs against motifs obtained from the RNAcompete assay.

4.1.2 Example of predicted Z-scores

(Fig. 4.1) shows an example of predicted versus experimental 8-mer Z-scores for an Oikopleura dioica homeodomain SNAPOd2T00005194001 assayed by Weirauch et al. [15]. The overall rank correlation of predicted and experimental Z-scores is high (Spearman ρ = .765), and 48% of the top 100 8-mers based on predicted Z-scores overlap with the top 100 8-mers determined from experimental Z-scores. Moreover, running the PWM-Align-Z algorithm on top 100 predicted 8-mers produces a motif similar to the one obtained from the top experimental 8-mers.

4.1.3 Motif visualization and comparison of affinity regression motifs with nearest neighbor motifs

For the RBP data set, using 10-fold cross-validation, we trained a Z-score model using 7-mer Z-scores (as reported in Ray et al. (2013) [14]), predicted the top 100 7-mers for held-out RBPs, and used these as input to our motif prediction algorithm PWM-Align-Z. For the RNA motif prediction, we used PWM-Align-Z to produce a position weight matrix (PWM) for each RBP RNAcompete experiment. We used the top 100 7-mers with highest predicted Z-scores as input to PWM-Align-Z to generate binding motifs. We found the choice of 7-mer Z-

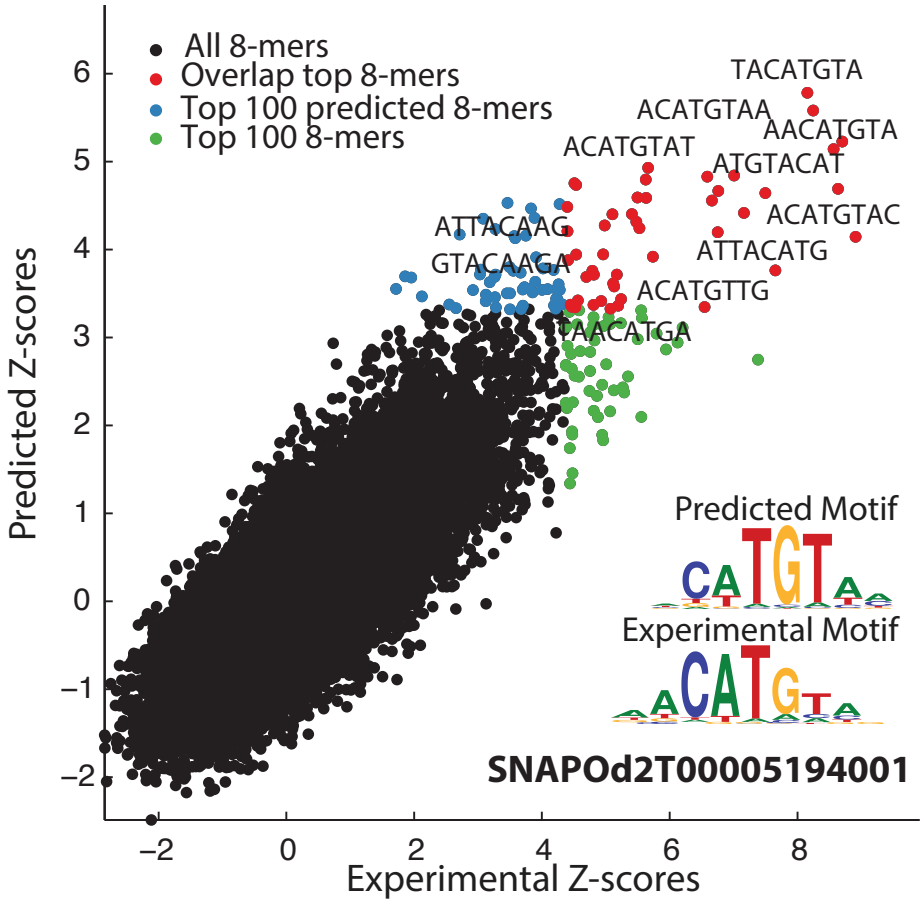

Figure 4.1: **Example of homeodomain predicted versus experimental Z-scores.** Example of predicted Z-scores from the Z-score affinity regression model, trained on 75 non-redundant mouse homeodomains, versus experimental Z-scores for SNAPOd2T00005194001, one of the diverse homeodomains assayed by Weirauch et al. [15] Binding motifs generated by PWM-Align-Z based on the top 100 8-mers predicted by affinity regression and the top 100 8-mers based on actual Z-scores are shown.

scores and top 100 7-mers for motif summarization to generate reproducible and reasonably high information content motifs across replicates. We visualized the PWMs from 207 RBPs, including both RRM and KH subfamilies using the motifStack (version 1.4.0) R package and plotted them in a circularized phylogenetic tree (Fig. 4.2).

4.1.4 Comparison of Kullback-Leibler divergence of affinity regression motifs with nearest neighbor motifs

I calculated the Kullback-Leibler divergence (D_{KL}) between the motifs generated from affinity regression and from the nearest neighbor algorithm and the ground truth motifs [40]. For each RBP we took its motif generated from affinity regression or nearest neighbor and generated from the RNA compete assay and calculated the minimum Kullback-Leibler divergence between the two motifs. We plotted the log D_{KL} - min log(D_{KL}) of the nearest neighbor motifs (x-axis) against the affinity regression motifs (y-axis) to compare the divergences from the ground truth motifs (Fig. 4.3). We plotted the probability density (Fig. 4.4) and cumulative distribution (Fig. 4.5) functions of log D_{KL} - min log(D_{KL}) for both affinity regression motifs and nearest neighbor motifs and see that the affinity regression motifs have a distribution of Kullback-Leibler divergences which is shifted to the left of the distribution for the nearest neighbor. In 10-fold crossvalidation on the full data set of RBPs, we found that the AR-predicted motifs were significantly closer to ground truth motifs (generated by the same motif algorithm on the experimental Z-scores) than nearest neighbor motifs (Wilcoxon signed rank test, $p < 7.66e\text{-}10$).

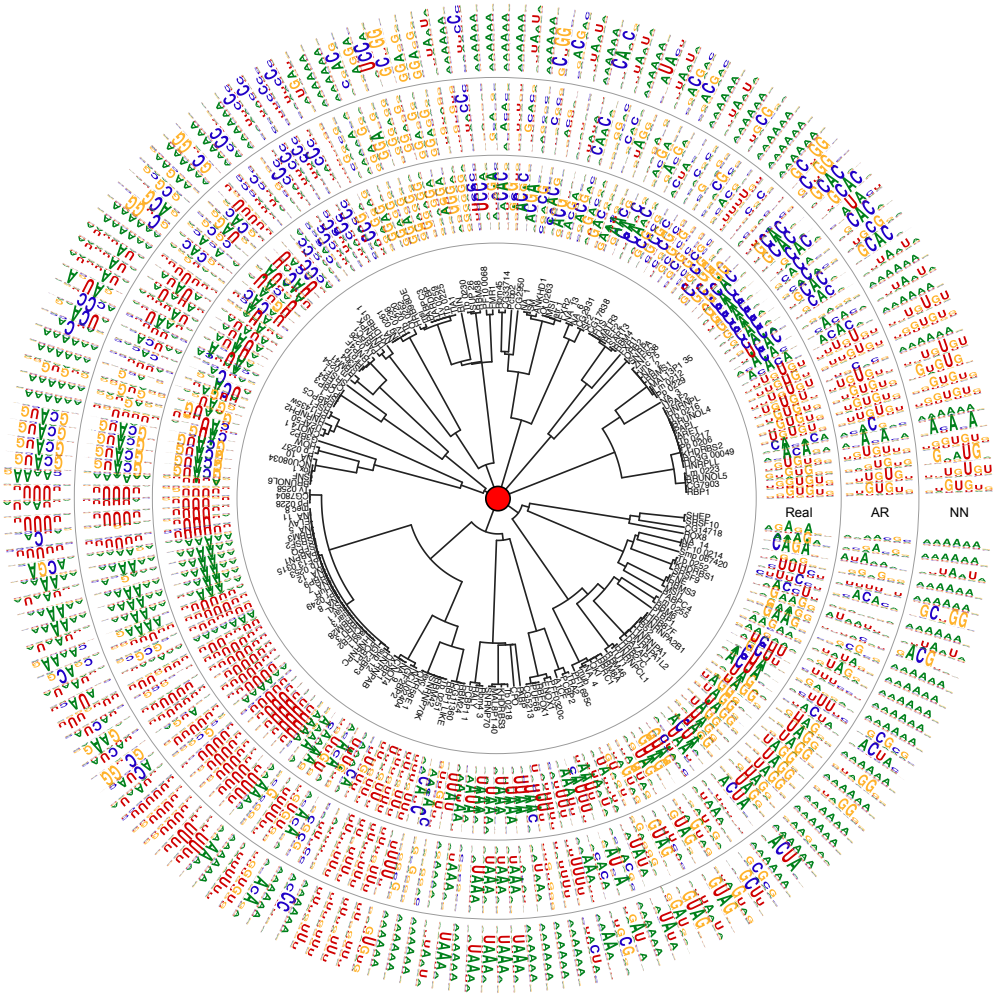

Figure 4.2: **AR-derived motif prediction for RBPs.** AR motifs are generated by running PWM-Align-Z on the top 100 7-mers as predicted for held-out RBPs using the Z-score affinity regression model (10-fold cross-validation). In the inner circle, we show the ground truth motif obtained from the experimental data Y, the middle circle shows motif obtained by AR, and the outer circle shows the motif obtained by NN. Plotted are predictions for both RRM and KH-I domains. The RRM motifs are well predicted by both AR and NN; KH family proteins are less well represented in the data set, and KH-I motifs are harder to predict for both methods.

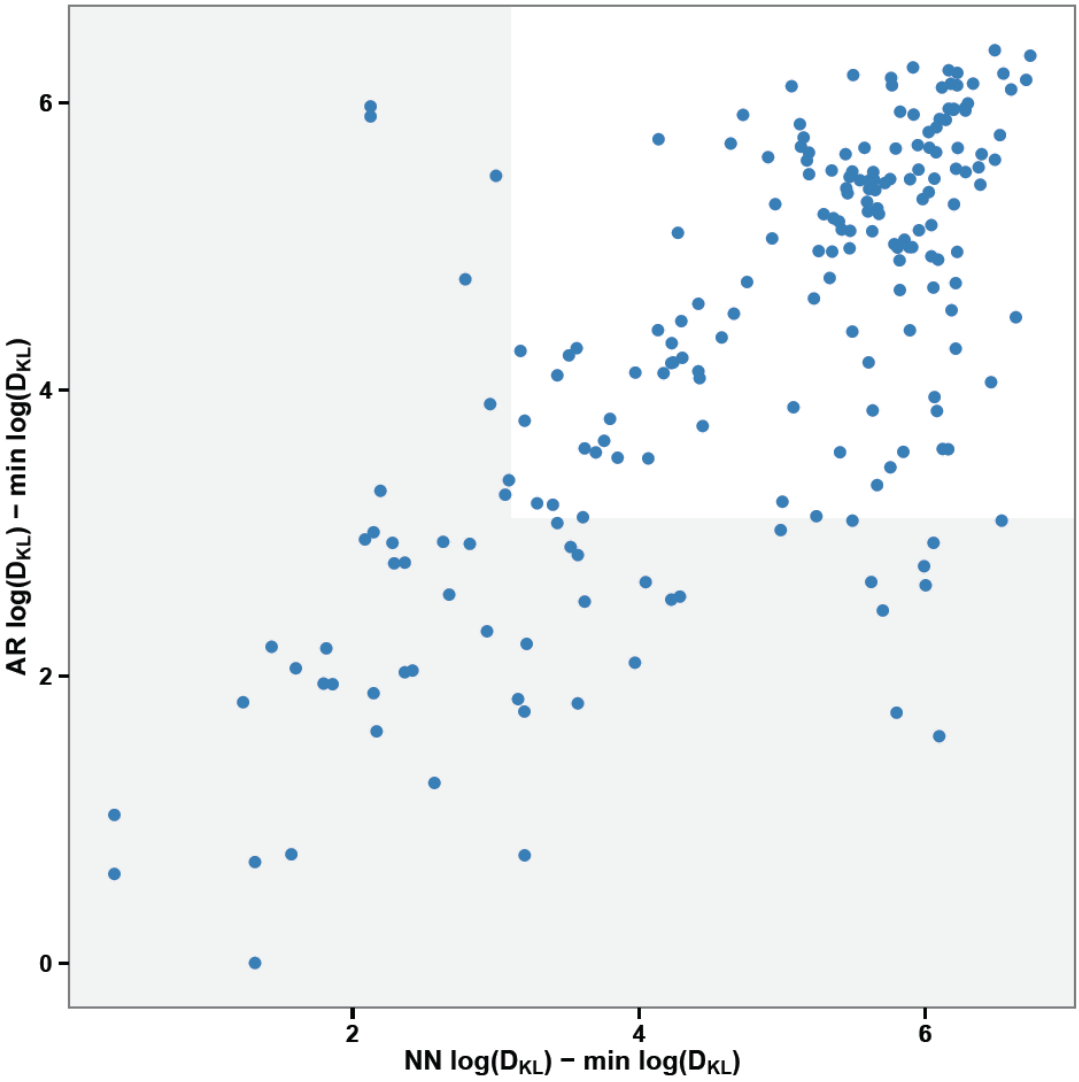

Figure 4.3: **RBP motif accuracy.** Predicted motifs were assessed for quality relative to ground truth by D_{KL}, computed by sliding the predicted PWM over the ground truth PWM and reporting the minimum D_{KL}. The NN (y-axis) and AR (x-axis) $\log(D_{KL})$ scores are plotted after subtracting the minimum $\log(D_{KL})$ for the data set. Motifs falling in the gray area satisfy a quality threshold equal to the median divergence between motifs from experimental replicates. AR-predicted motifs are significantly more accurate than NN motifs ($p < 7.66 \times 10\text{-}10$, Wilcoxon signed rank test).

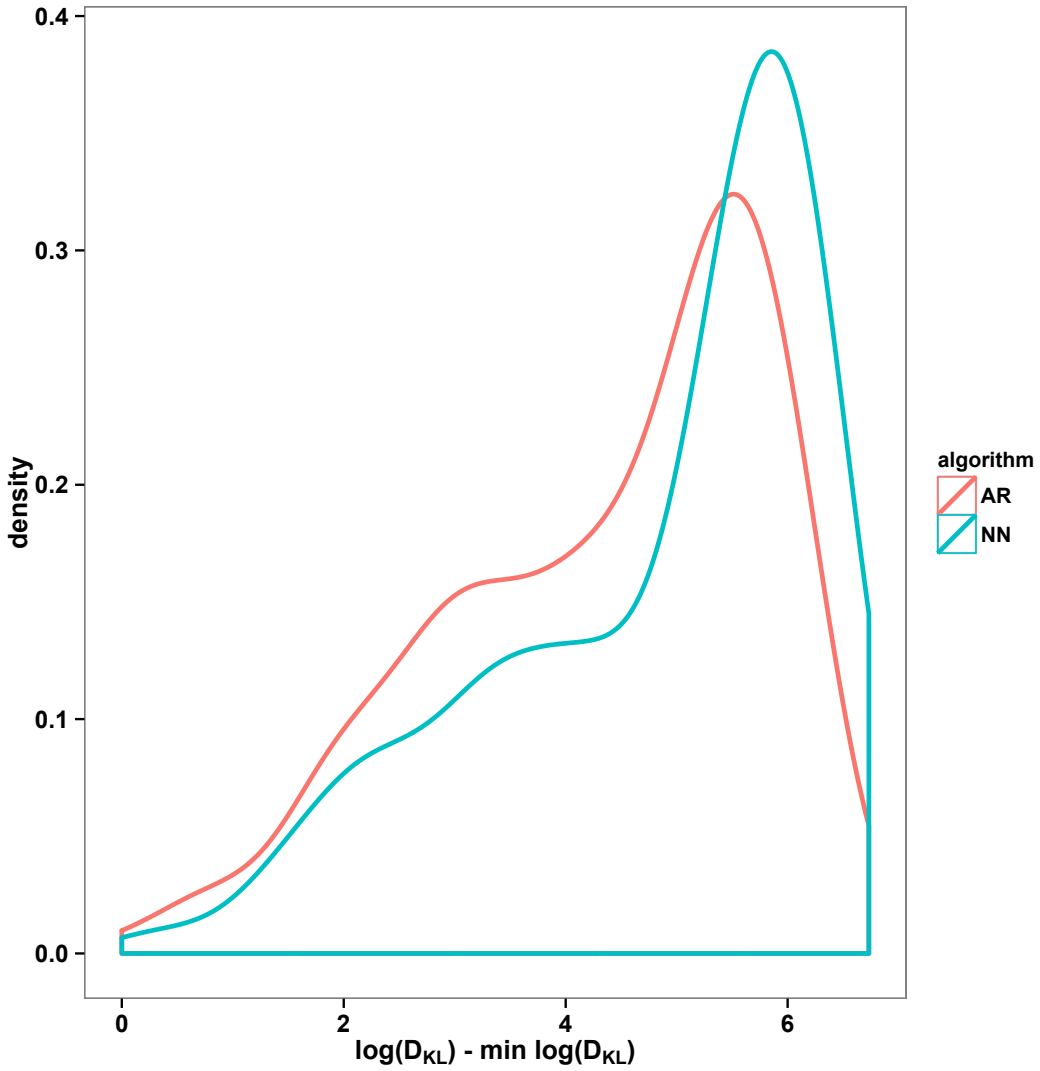

Figure 4.4: **Probability density function of $\log(D_{KL})$ for affinity regression and nearest neighbor.** Probability density function of held-out $\log(D_{KL})$ for AR and NN.

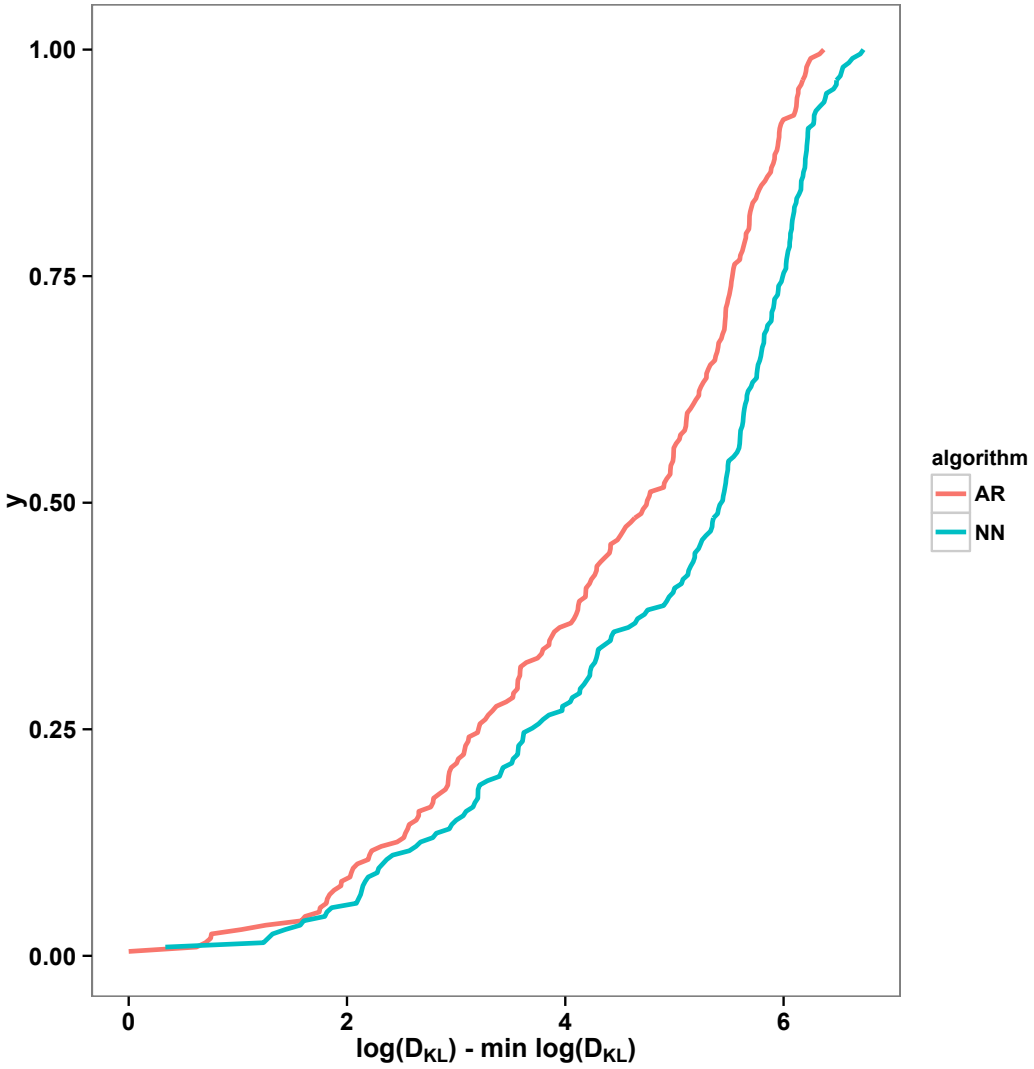

Figure 4.5: **Cumulative distribution function of** $\log(D_{KL})$ **for affinity regression and nearest neighbor.** Cumulative distribution function of held-out $\log(D_{KL})$ for AR and NN.

CHAPTER 5

CONCLUSION

5.1 Conclusion and future directions

5.1.1 Introduction

In recent years researcher have found evidence of microenvironment-mediated drug resistance to targeted cancer therapies [2, 9, 38, 45, 47], recently generating two high throughput data sets studying stromal mediated drug resistance in co-cultures [41, 48]. In my dissertation, I developed a new supervised learning strategy for modeling cancer-stromal cell paracrine signaling from anti-cancer drug sensitivity co-culture experiments. In the co-culture analysis, our modeling was limited by the size of our data set. In this context, our modeling provides a proof-of-principle analysis that could be applied to larger co-culture data sets in the future and that could yield novel insights into interactions mediating drug resistance.

In this dissertation we applied supervised learning methods with pairwise features including multi-task affinity regression and multi-task pairwise SVM, to model biological interactions in two different contexts. In one context we modeled microenvironment-mediated drug resistance to targeted cancer therapies. In a second context, we modeled the binding affinity of nucleic acid and proteins in PBM and RNAcompete experiments. This dissertation is intended to serve as both (1) a point of reference for experimental biologists looking to use practical computational methods to model tumor-stromal interactions in stud-

ies of the molecular mechanisms mediating innate drug resistance and (2) more generally a starting point for computational biologists interested in modeling biological interaction data.

5.1.2 Stromal-mediated drug resistance

In this dissertation, I explored modeling microenvironment-mediated drug resistance as the interaction between molecular components of cancer cells and stromal cells [30]. Using an expression-based pathway feature representation for cancer cells and a cytokine expression level representation for stromal cells, I trained affinity regression models to predict stromal-mediated rescue scores of cancer cells for each drug; I employed a multi-task strategy to share information across models for different dosages of the same drug or different drugs of the same class. As a bilinear regression model, affinity regression allows us to define feature space mappings using the trained model. The mappings identified the cytokines that are most strongly associated with resistance/sensitivity for each cancer cell line, as well as the cancer cell pathways that appear to mediate resistance/sensitivity for each stromal cell line. Through an empirical null model, I assigned statistical significance to these key predicted features. This analysis recovered the finding that stromal-derived HGF mediates resistance to BRAF inhibitors in melanoma cell lines. Moreover, our affinity regression analysis predicted that HGF would mediate resistance to EGFR inhibitors in lung cancer cell lines. Follow-up experiments with afatinib and erlonitib confirmed this clinically relevant prediction. I also explored an alternative supervised learning method that models biological interaction data using a multi-task pairwise SVM. I compare the performance of the SVM with affinity regression using 10-

fold cross-validation trained on a tumor-stromal co-culture drug screen data set.

The largest limitation of our study is the small training data set size. Ideally, we would use on the order of 100 stromal and cancer cell lines with a full matrix of co-culture experiments as training data for each drug and dosage; in practice, the co-culture matrix consisted of 10-20 cancer cell lines by a similar number of stromal cell lines, and a different matrix of co-cultures was assayed for each drug. While our multi-task strategy helps improve model accuracy in this low training data setting, it cannot fully address the fact that we are not adequately sampling the space of cancer-stromal interactions. The current work provides a proof-of-principle that supervised learning can derive novel findings from co-culture drug sensitivity data sets, providing a path forward for future larger-scale studies.

Another limitation of our study was the fact that we were using available static data sets from the cancer and stromal cell lines in monoculture. On the cancer side, we had gene expression data collected from cancer cell lines in monoculture. On the stromal side we had cytokine expression collected from stromal cell secretion in monoculture. If we had gene expression data collected from cancer cell lines in co-culture experiments and if we had stromal cell cytokine secretion from co-culture data, it would give more dynamic rather than static sources of information for our modeling. This could potentially yield novel insights into the paracrine signaling mechanisms between tumor and stromal cells.

In recent years, two high-throughput screens have produced a significant amount of data that has added to our understanding of resistance to molecularly targeted therapies [41, 48]. The identification of HGF as a mediator of

vemurafenib resistance has had translational impact. Both groups suggest clinical investigation of the concurrent treatment of BRAF inhibitors with the FDA-approved MET inhibitor crizotinib. The work in this dissertation to understand drug resistance against targeted therapies mediated by the tumor microenvironment could potentially lead to combination drug therapies targeting both the tumor and the microenvironment. Therapies that target oncogenic signaling pathways of cancer cells while manipulating secreted factors of the microenvironment may provide a new treatment option to abrogate cancer resistance in clinical settings.

The genomics field has become increasingly data rich due to the availability of high-throughput technologies such as next-generation sequencing and microarray profiling. In our analysis we made use of the mRNA expression signature representation of cancer cell lines due to the availability of large cell line profiling studies. Looking forward, it is imaginable that high-throughput techniques such as RNA-Seq, ChIP-Seq, miRNA chips, lncRNA profiling, and next-generation sequencing can produce a large pool of data that can be used to identify predictive biomarkers to personalize patient treatment. For our analysis, phosphoproteomic data might be more directly relevant as a representation of active signaling pathways in cancer cells. Data resources based on technologies like reverse-phase protein array [25] and mass spectrometry [31] are growing. However, the overlap of available proteomics data sets with the cancer cell lines in our study remains limited. Perhaps once new data sources become available in the future, they could be leveraged using the supervised learning methods proposed in this dissertation.

5.1.3 Learning families of transcription factors and RNA binding proteins from PBM and RNAcompete data

In another context, affinity regression was used to model nucleic acid-protein binding in PBM and RNAcompete experiments. In this context, we used affinity regression to learn the DNA or RNA recognition code for families of TFs or RBPs directly from the protein sequence and probe-level binding data from PBM or RNAcompete experiments. Affinity regression trains on PBM or RNAcompete experiments to learn an interaction model between proteins and nucleic acids. Learning from PBM or RNAcompete data for diverse TFs and RBPs, the affinity regression model can predict the binding affinities of held-out proteins and identify key DNA/RNA-binding residues. Affinity regression recovers predicted Z-scores with high correlation with experimental Z-scores and with high overlap between the top 8-mers from the predicted and experimental Z-scores. Affinity regression recovered motifs with higher motif accuracy or smaller Kullback-Leibler divergence to ground truth motifs than nearest neighbor motifs.

Future directions for learning families of transcription factors and RNA binding proteins from PBM and RNAcompete data

Our affinity regression model can be extended to predict binding affinity of new proteins in future PBM and RNAcompete experiments. In this study we predicted the binding affinity of homeodomain proteins. The affinity regression method can also be applied to other transcription factor families whose binding affinities to DNA probes were measured in PBM experiments [34]. Other tran-

scription factor families whose binding affinity could be modeled with affinity regression include: C2H2 ZF, bZIP, Zinc cluster, Myb/SANT, bHLH, Nuclear receptor, AP2, GATA, Sox, and Forkhead transcription factor families. It would be interesting to look at the predicted DNA binding residues of the transcription factor families other than homeodomain and compare their binding residues to the residues of the homeodomain family.

5.1.4 Conclusion

In my dissertation, I applied two supervised learning methods: multi-task affinity regression and multi-task pairwise SVM to cell co-culture systems with quantitative phenotypes. Affinity regression was also used to learn the binding recognition code for families of TFs and RBPs from PBM/RNAcompete data. The DNA recognition code for a family of TFs or RBPs was learned from PBM or RNAcompete data. More broadly, the affinity regression can be used to train a bilinear interaction model for any macromolecular or cellular interactions where interactors are described by features and where a high-throughput affinity readout is available and these supervised learning methods may be used as a general strategy to model and interpret biological interaction data.

BIBLIOGRAPHY

[1] Prahallad A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, Beijersbergen RL, Bardelli A, and Bernards R. **Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR**. *Nature*, 483(7687):100–3, 2012.

[2] Obenauf AC, Zou Y, Ji AL, Vanharanta S, Shu W, Shi H, Kong X, Bosenberg MC, Wiesner T, Rosen N, Lo RS, and Massague J. **Therapy-induced tumour secretomes promote resistance and tumour progression**. *Nature*, 520(7547):368–372, 2015.

[3] Brunner C, Fischer A, Luig K, and Thies T. **Pairwise Support Vector Machines and their Application to Large Scale Problems**. *Journal of Machine Learning*, 13:2279–2292, 2012.

[4] Montero-Conde C, Ruiz-Llorente S, Dominguez JM, Knauf JA, Viale A, Sherman EJ, Ryder M, Ghossein RA, Rosen N, and Fagin JA. **Relief of Feedback Inhibition of HER3 Transcription by RAF and MEK Inhibitors Attenuates Their Antitumor Effects in BRAF-Mutant Thyroid Carcinomas**. *Cancer Discovery*, 3(5):520–533, 2013.

[5] Britten CD. **Targeting ErbB receptor signaling: A pan-ErbB approach to cancer**. *Molecular Cancer Therapeutics*, 3(10):1335–42, 2004.

[6] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, and et al. **PID: the Pathway Interaction Database**. *Nucleic Acids Res*, 37:674–9, 2009.

[7] Bixby D and Talpaz M. **Seeking the causes and solutions to imatinib-resistance in chronic myeloid leukemia**. *Leukemia*, 25:722, 2011.

[8] Quail DF and Joyce JA. **Microenvironmental regulation of tumor progression and metastasis**. *Nature Medicine*, 19(11):1423–37, 2013.

[9] DeNardo DG, Brennan DJ, Rexhepaj E, Ruffell B, Shiao SL, Madden SF, Gallagher WM, Wadhwani N, Keil SD, Junaid SA, Rugo HS, Hwang ES, Jirstrm K, West BL, and Coussens LM. **Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy**. *Cancer Discov*, 1(1):54–67, 2011.

[10] Bottaro DP, Rubin JS, Faletto DL, Chan AM, Kmiecik TE, Vande Woude GF,

and et al. **Identification of the hepatocyte growth factor receptor as the c-met proto-oncogene product**. *Science*, 251(4995):802–4, 1991.

[11] Anderson DR, Grillo-Lpez A, Varns C, Chambers KS, and Hanna N. **Targeted anti-cancer therapy using rituximab, a chimaeric anti-CD20 antibody (IDEC-C2B8) in the treatment of non-Hodgkin's B-cell lymphoma**. *Biochem Soc Trans*, 25(2):705–8, 1997.

[12] Berger MF et al. **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nature Biotechnology*, 24:14291435, 2006.

[13] Ray D et al. **Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins**. *Nature Biotechnology*, 27:667670, 2009.

[14] Ray D et al. **A compendium of RNA-binding motifs for decoding gene regulation**. *Nature*, 499:172177, 2013.

[15] Weirauch MT et al. **Determination and inference of eukaryotic transcription factor sequence specificity**. *Cell*, 158:14311443, 2014.

[16] Klemm F and Joyce JA. **Microenvironmental regulation of therapeutic response in cancer**. *Trends Cell Biol*, 25(4):198–213, 2015.

[17] Hotelling H. **Relations Between Two Sets of Variates**. *Biometrika*, 28(3-4):321377, 1936.

[18] Wold H. **Estimation of principal components and related models by iterative least squares**. *Krishnaiaah, P.R. Multivariate Analysis*, 243:391–420, 1966.

[19] Zahreddine H and Borden KLB. **Mechanisms and insights into drug resistance in cancer**. *Frontiers in Pharmacology*, 4(28), 2013.

[20] Kantarjian HM and Talpaz M. **Imatinib mesylate: clinical results in Philadelphia chromosome-positive leukemias**. *Semin Oncol*, 28(5):9–18, 2001.

[21] Osmanbeyoglu HU, Pelossof R, Bromberg JF, and Leslie C. **Linking signaling pathways to transcriptional programs in breast cancer**. *Genome Res*, 24(11):1869–80, 2014.

[22] Satzger I, Kuttler U, Volker B, Schenck F, Kapp A, and Gutzmer R. **Anal mucosal melanoma with KIT-activating mutation and response to imatinib therapycase report and review of the literature**. *Dermatology*, 220(1):7781, 2010.

[23] Albanell J and Baselga J. **Trastuzumab, a humanized anti-HER2 monoclonal antibody, for the treatment of breast cancer**. *Drugs Today*, 35(12):931–46, 1999.

[24] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, and et al. **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature*, 483(7391):603–7, 2012.

[25] Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, and et al. **TCPA: a resource for cancer functional proteomics data**. *Nat Methods*, 10(11):1046–7, 2013.

[26] Joyce JA and Pollard JW. **Microenvironmental regulation of metastasis**. *Nature Reviews Cancer*, 9:239–252, 2009.

[27] Sosman JA, Kim KB, Schuchter L, Gonzalez R, Pavlick AC, Weber JS, McArthur GA, HutsonTE, Moschos SJ, Flaherty KT, Hersey P, Kefford R, Lawrenc D, Puzanov I, Lewis KD, Amaravadi RK, Chmielowski B, Lawrence HJ, Shyr Y, Ye F, Li J, Nolop KB, Lee RJ, Joe AK, and Ribas A. **Survival in BRAF V600-mutant advanced melanoma treated with Vemurafenib**. *N. Engl. J. Med.*, 366:707–714, 2012.

[28] Wilken JA and Maihle NJ. **Primary trastuzumab resistance: new tricks for an old drug**. *Annals of the New York Academy of Sciences*, 10:1749–6632, 2010.

[29] Rowley JD. **A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining**. *Nature*, 243:290–293, 1973.

[30] Yang JL, Bowman RL, Pelossof R, Joyce JA, and Leslie CS. **Modeling tumor-stromal interactions that mediate innate resistance to targeted cancer therapies**. *In process of submission*, 2016.

[31] Whiteaker JR, Halusa GN, Hoofnagle AN, Sharma V, MacLean B, Yan P, and et al. **CPTAC Assay Portal: a repository of targeted proteomic assays**. *Nat Methods*, 11(7):703–4, 2014.

[32] Ruben KG. **Generalised Bilinear Regression**. *Biometrika*, 85(3):689–700, 1998.

[33] Flaherty KT, Puzanov I, and Kim KB. **Inhibition of mutated, activated BRAF in metastatic melanoma**. *New England Journal of Medicine*, 363:809819, 2010.

[34] Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, and Hughes TR. **Determination and inference of eukaryotic transcription factor sequence specificity**. *Cell*, 158(6):1431–1443, 2014.

[35] Lito P, Pratilas CA, Joseph EW, Tadi M, Halilovic E, Zubrowski M, Huang A, Wong WL, Callahan MK, Merghoub T, Wolchok JD, de Stanchina E, Chandarlapaty S, Poulikakos PI, Fagin JA, and Rosen N. **Relief of profound feedback inhibition of mitogenic signaling by RAF inhibitors attenuates their activity in BRAFV600E melanomas**. *Cancer Cell*, 22(5):668–682, 2012.

[36] Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, Shi H, Atefi M, Titz B, Gabay MT, Salton M, Dahlman KB, Tadi M Wargo JA, Flaherty KT, Kelley MC, Misteli T, Chapman PB, Sosman JA, Graeber TG, Ribas A, Lo RS, Rosen N, and Solit DB. **RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E)**. *Nature*, 480:387–390, 2011.

[37] Dummer R and Flaherty KT. **Resistance patterns with tyrosine kinase inhibitors in melanoma: new insights**. *Curr Opin Oncol*, 24:150–154, 2012.

[38] Hughes R, Qian BZ, Rowan C, Muthana M, Keklikoglou I, Olson OC, Tazzyman S, Danson S, Addison C, Clemons M, Gonzalez-Angulo AM, Joyce JA, De Palma M, Pollard JW, and Lewis CE. **Perivascular M2 Macrophages Stimulate Tumor Relapse after Chemotherapy**. *Cancer Res*, 75(17):3479–3491, 2015.

[39] Nazarian R, Shi H, Wang Q, Kong X, Koya RC, Lee H, Chen Z, Lee M, Attar N, Sazegar H, Chodon T, Nelson SF, McArthur G, Sosman JA, Ribas A, and Lo RS. **Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation**. *Nature*, 468(7326):9737, 2010.

[40] Pelossof R, Singh I, Yang JL, Weirauch MT, Hughes TR, and Leslie CS. **Affinity regression predicts the recognition code of nucleic acid-binding proteins**. *Nature Biotechnology*, 33(12):1242–1249, 2015.

[41] Straussman R, Morikawa T, Shee K, Barzily-Rokni M, Qian ZR, Du J, Davis A, Mongare MM, Gould J, Frederick DT, Cooper ZA, Chapman PB, Solit DB, Ribas A, Lo RS, Flaherty KT, Ogino S, Wargo JA, and Golub TR. **Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion**. *Nature*, 487(7408):500–504, 2012.

[42] Wadlow RC, Wittner BS, Finley SA, Bergquist H, Upadhyay R, Finn S, and et al. **Systems-level modeling of cancer-fibroblast interaction**. *PloS one*, 4(9):6888, 2009.

[43] Finn RS, Dering J, Conklin D, Kalous O, Cohen DJ, Desai AJ, and et al. **A selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro**. *Breast cancer research*, 11(5):77, 2009.

[44] Acharyya S, Oskarsson T, and Vanharanta S. **A CXCL1 paracrine network links cancer chemoresistance and metastasis**. *Cell*, 150:165–178, 2012.

[45] Acharyya S, Oskarsson T, Vanharanta S, Malladi S, Kim J, Morris PG, Manova-Todorova K, Leversha M, Hogg N, Seshan VE, Norton L, Brogi E, and Massague J. **A CXCL1 paracrine network links cancer chemoresistance and metastasis**. *Cell*, 150(1):165–178, 2012.

[46] Evgeniou T and Pontil M. **Regularized Multi-Task Learning**. *ACM*, 2004.

[47] HShree T, Olson OC, Elie BT, Kester JC, Garfall AL, Simpson K, Bell-McGuinn KM, Zabor EC, Brogi E, and Joyce JA. **Macrophages and cathepsin proteases blunt chemotherapeutic response in breast cancer**. *Genes Dev*, 25(23):2465–2479, 2011.

[48] Wilson TR, Fridlyand J, Yan Y, Penuel E, Burton L, Chan E, Peng J, Lin E, Wang Y, Sosman J, Ribas A, Li J, Moffat J, Sutherlin DP, Koeppen H, Merchant M, Neve R, and Settleman J. **Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors**. *Nature*, 487(7408):505–509, 2012.

[49] Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, and et al. **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for thera-**

peutic biomarker discovery in cancer cells. *Nucleic Acids Res*, 41:955–61, 2013.